Preprocessing Microarray Data: Beyond Expression

Rafael A. Irizarry

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

Acknowledgements

- Zhijin Wu, Brown University
- Benilton Carvalho, JHU
- Hao Wu, JHU
- Wenyi Wang, JHU
- Terry Speed, UC Berkeley

Outline

- Expression Arrays (15 minutes)

- SNP chips (15 minutes)

- Tiling Arrays (5 minutes)

Software: oligo package

Expression: Image -> Feature level -> Gene level

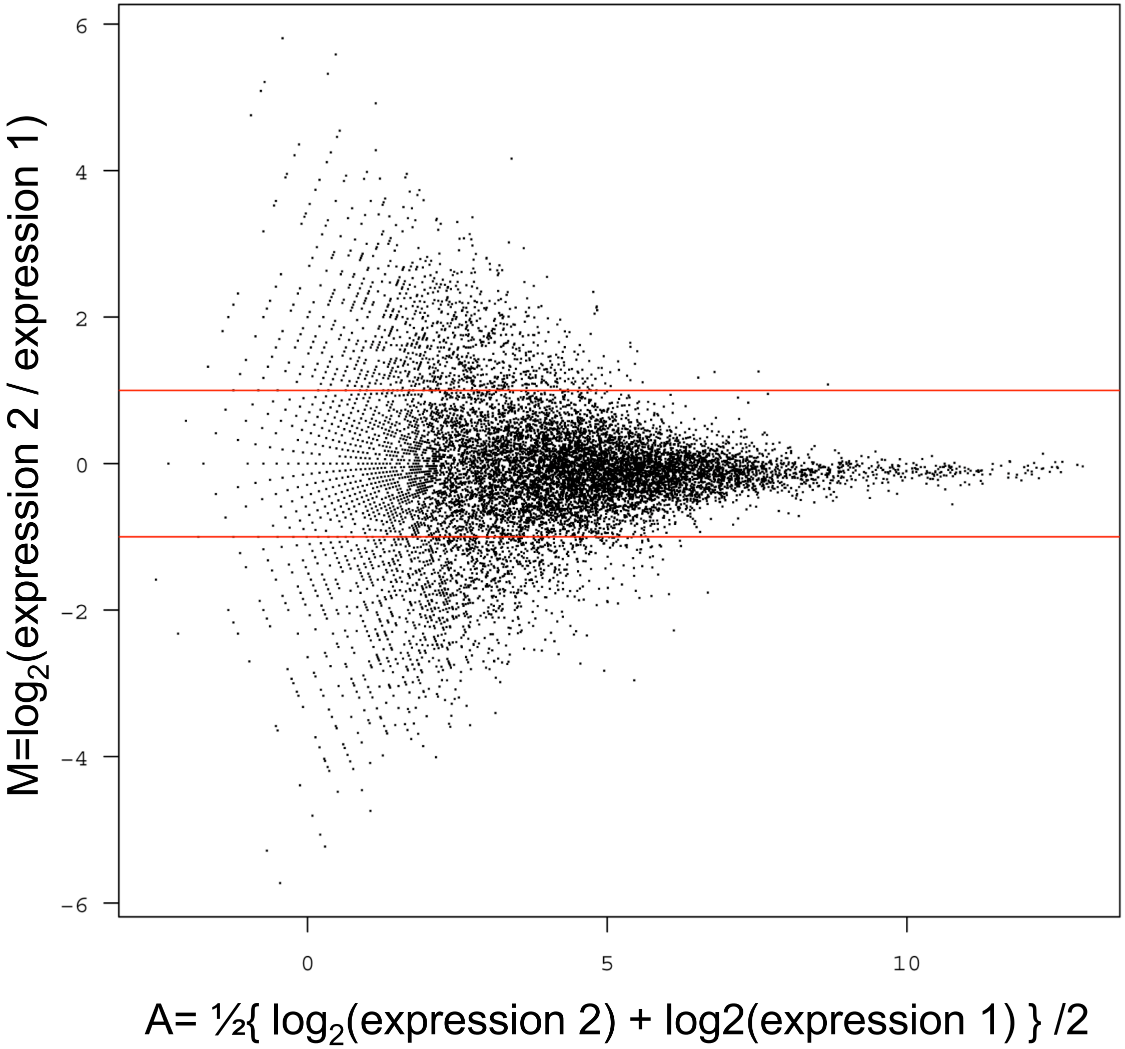
SNP: Image -> Feature level -> SNP Q level -> Call level

Tiling: Image -> Feature level -> ?

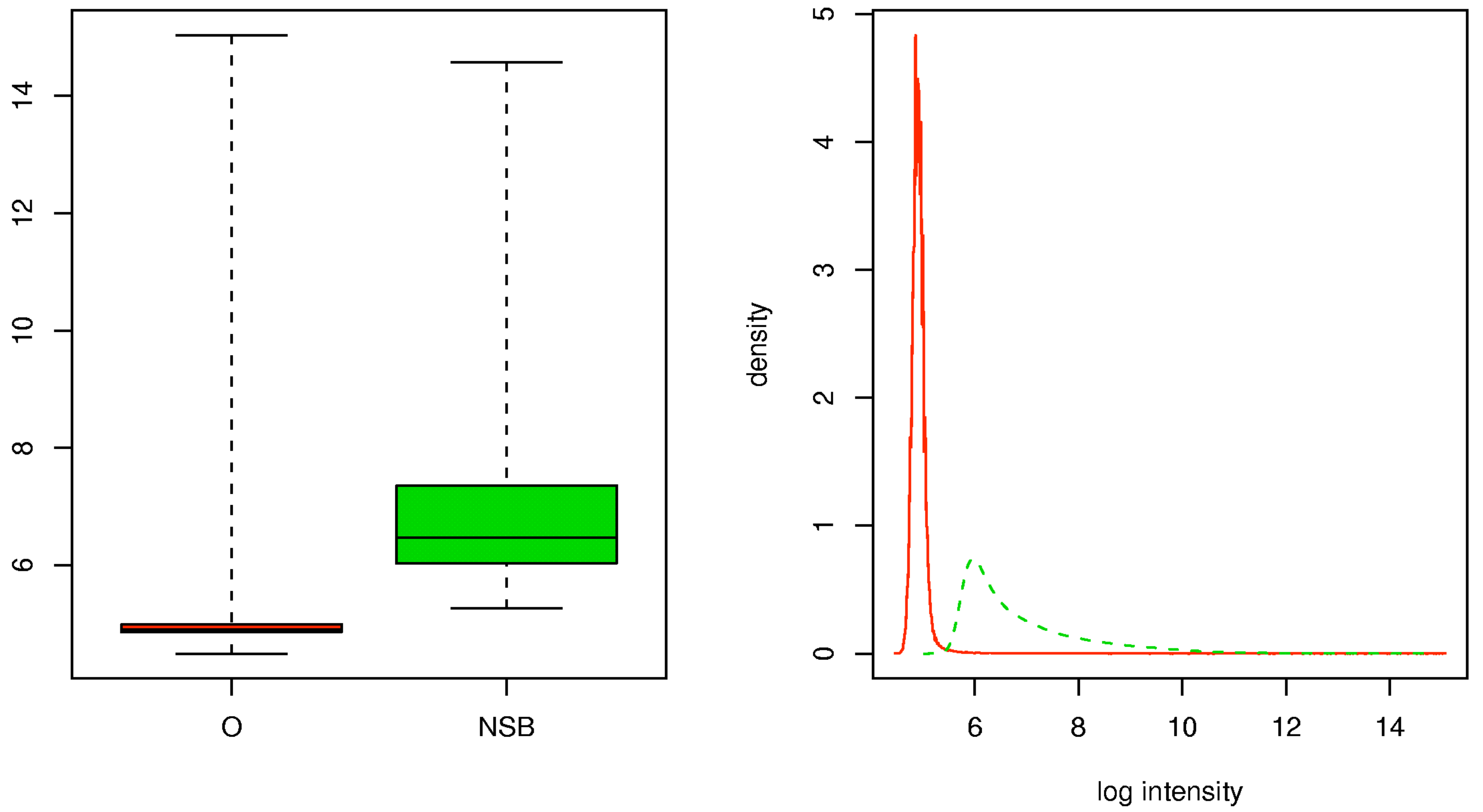
Common tasks: BG correction, Normalization, Sequence effects, Summarization

Expression Arrays

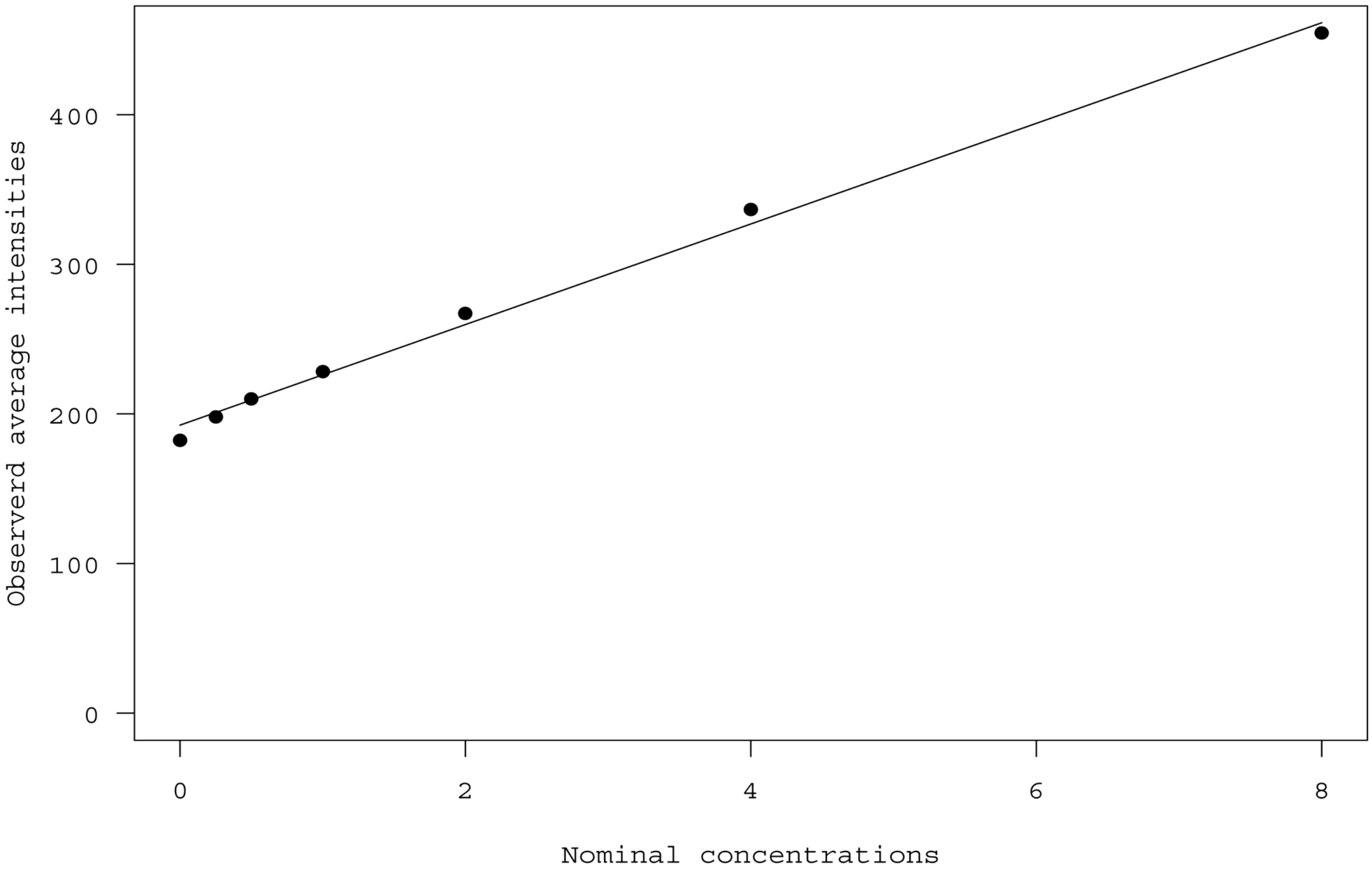
MvA Plot


M=log$_2$(expression 2 / expression 1)
6
4
2
0
−2
−4
−6
0
5
10
A= ½{ log$_2$(expression 2) + log2(expression 1) } /2

Background Noise

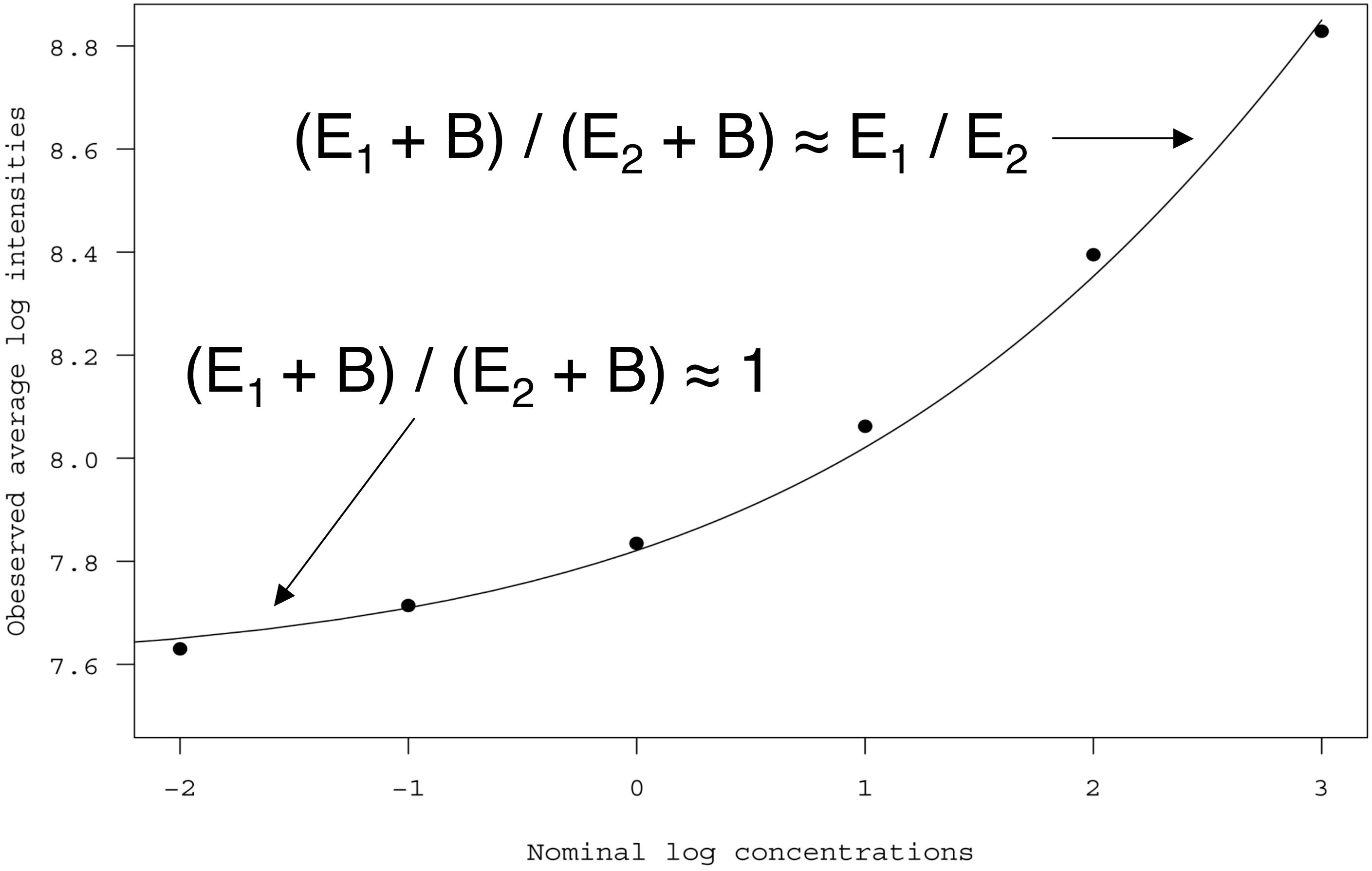


Why adjust?



Observerd average intensities
0
100
200
300
400
0
2
4
6
8
Nominal concentrations

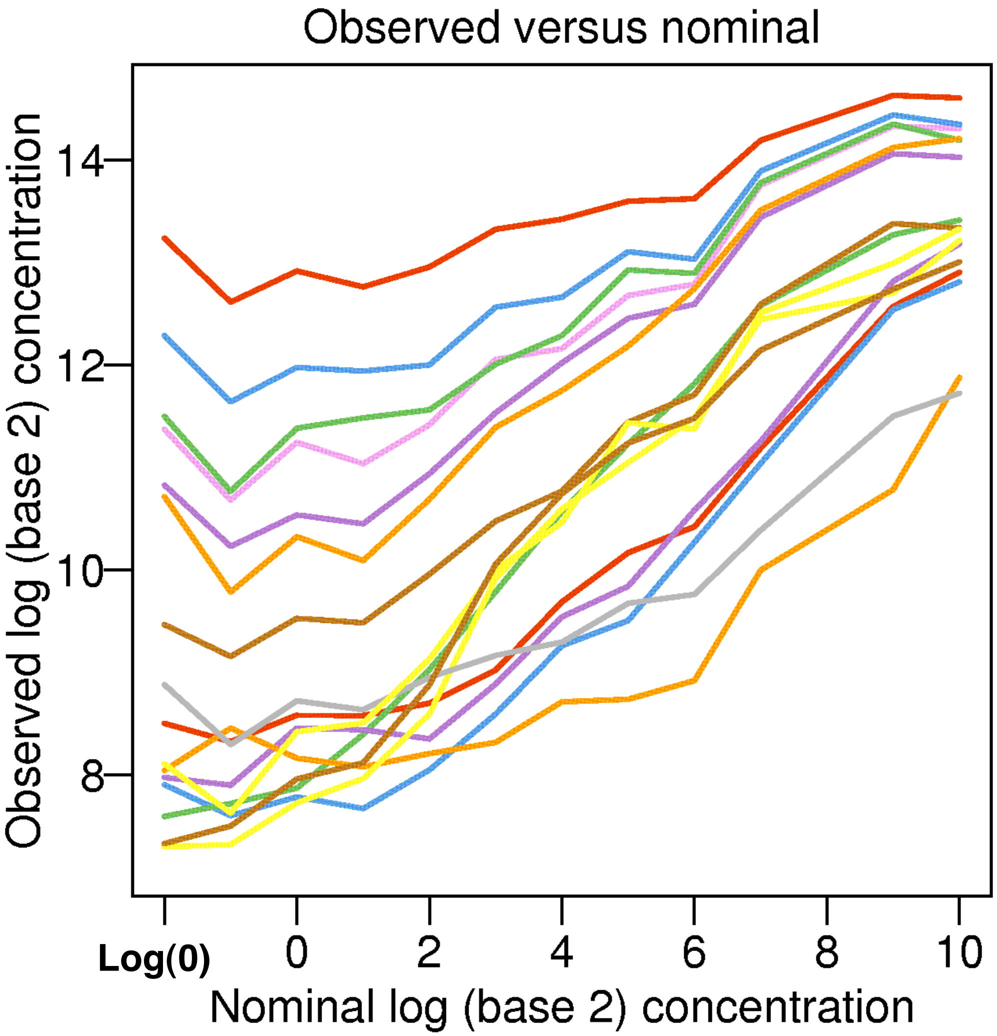
Why adjust?



$(E_1 + B) / (E_2 + B) \approx E_1 / E_2$

$(E_1 + B) / (E_2 + B) \approx 1$

Probe specific background

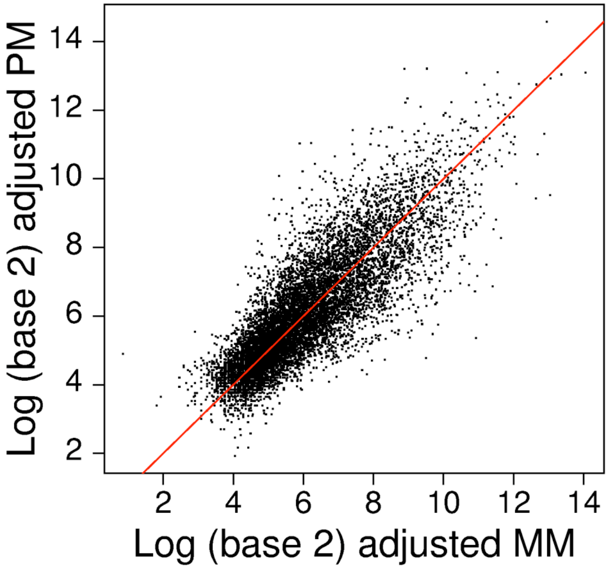


Direct Measurement Strategy

The hope is that:

$$PM = B + S$$
$$MM = B$$
$$\rightarrow$$
$$PM - MM = S$$

But this is not correct!



Notice

- **We care about ratios**
- **We usually take log of S**

Stochastic Model

Better to assume:

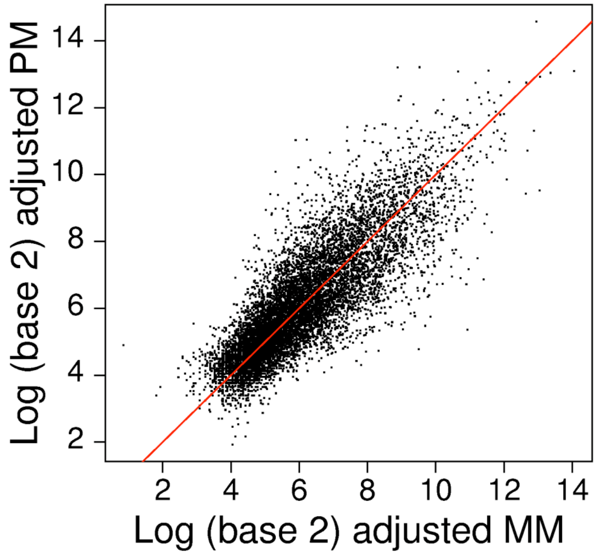
$$PM = B_{PM} + S$$

$$MM = B_{MM}$$

$$Cor[\log(B_{PM}), \log(B_{MM})\ ]=0.7$$

$$\Longrightarrow \quad Var[\log(PM - MM^{*})] \sim 1/S^{2}$$

Consider model based solutions and minimize MSE



General Model

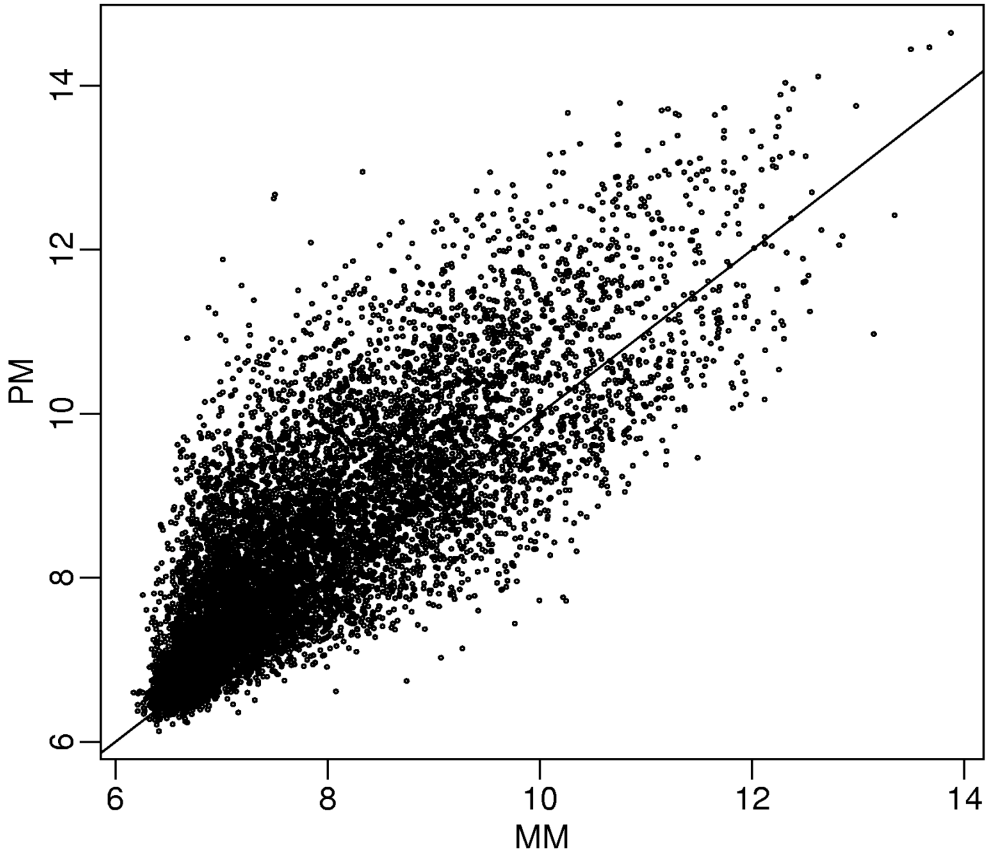
$$PM_{gij} = O_i^{PM} + \overbrace{\exp(h_i(\alpha_j^{PM}) + b_{gj}^{PM} + \varepsilon_{gij}^{PM})}^{\text{NSB}} + \overbrace{\exp(f_i(\alpha_j) + \theta_{gi} + \xi_{gij})}^{\text{SB}}$$

$$MM_{gij} = O_i^{MM} + \exp(h_i(\alpha_j^{MM}) + b_{gj}^{MM} + \varepsilon_{gij}^{MM})$$

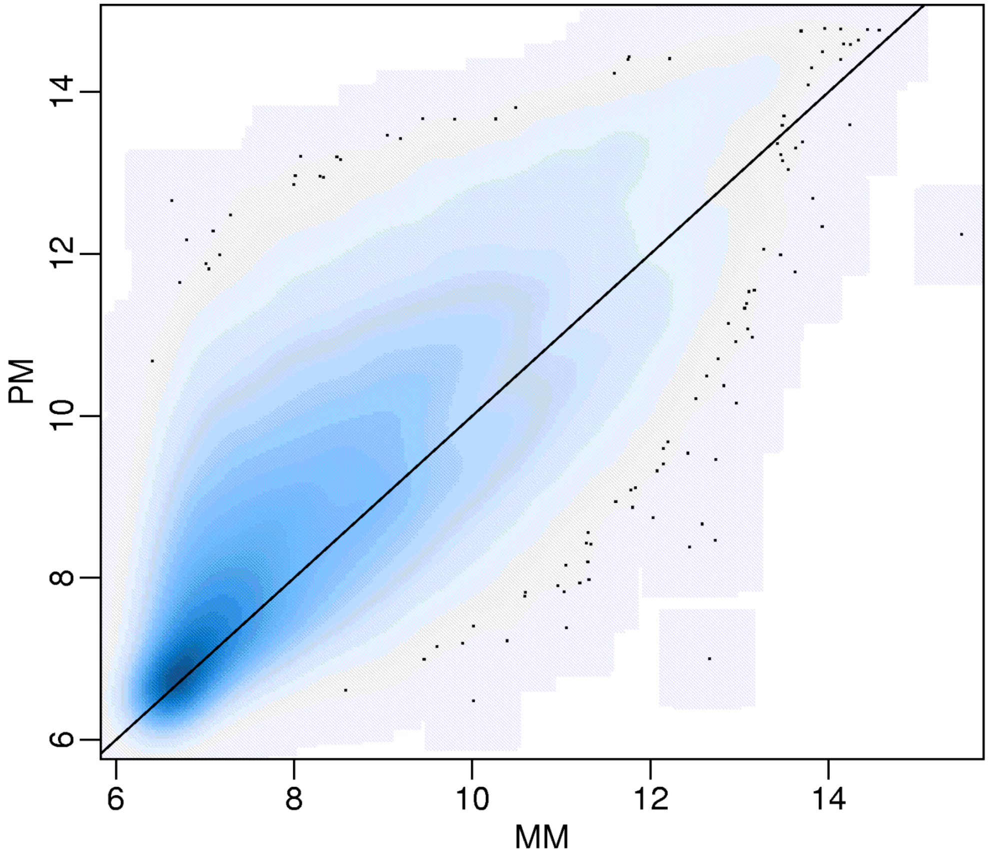
We can calculate: $E[T(\theta_g) \,|\, PM_g, MM_g]$

RMA uses a very simple model that provides a closed form version, ignores MM

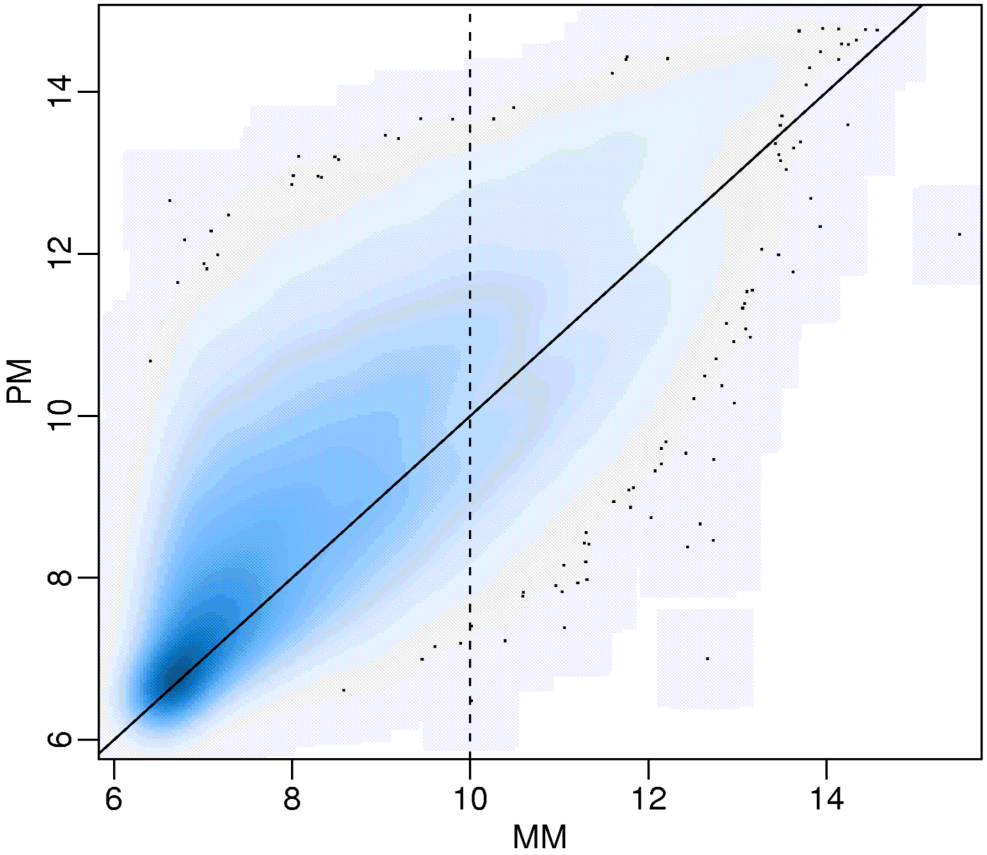
Why we did not use MM



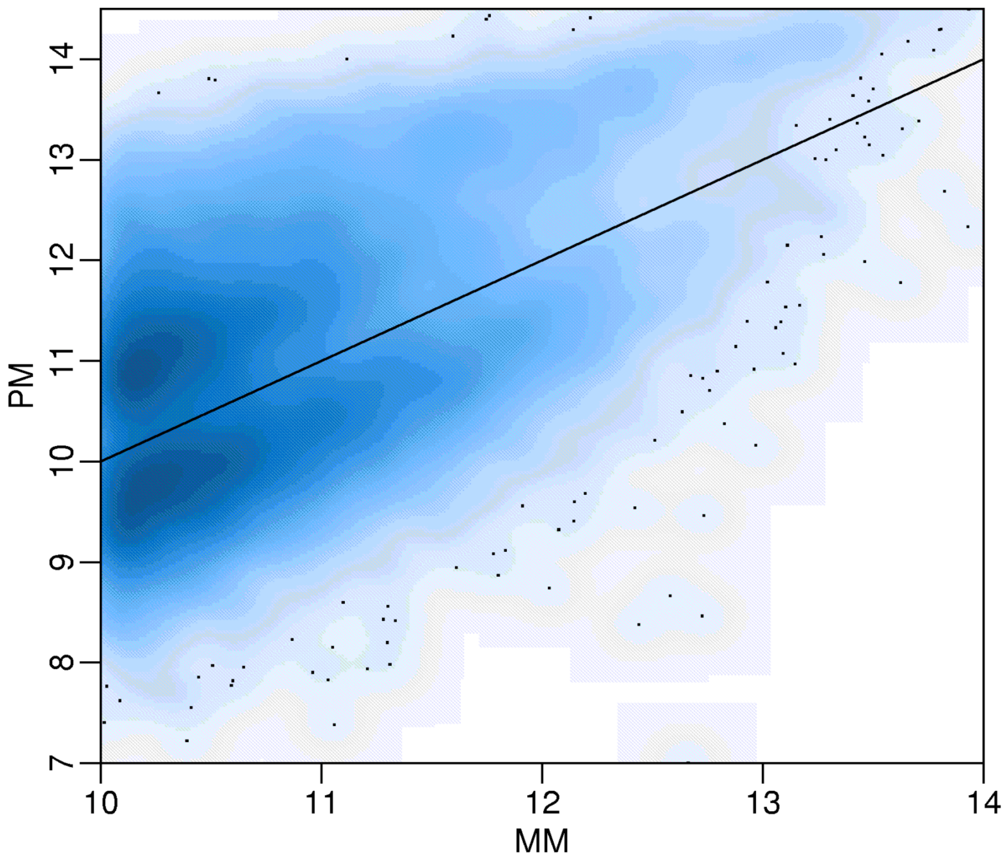
Two modes



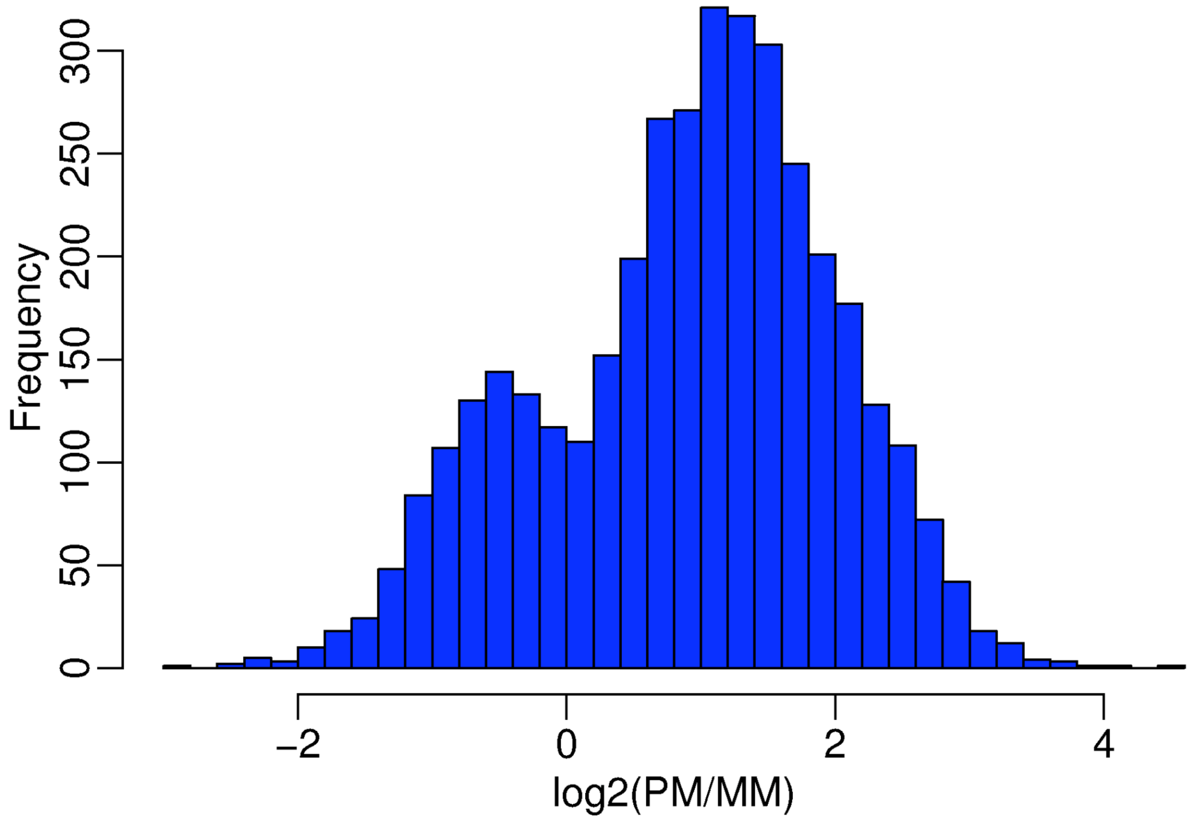
Two modes



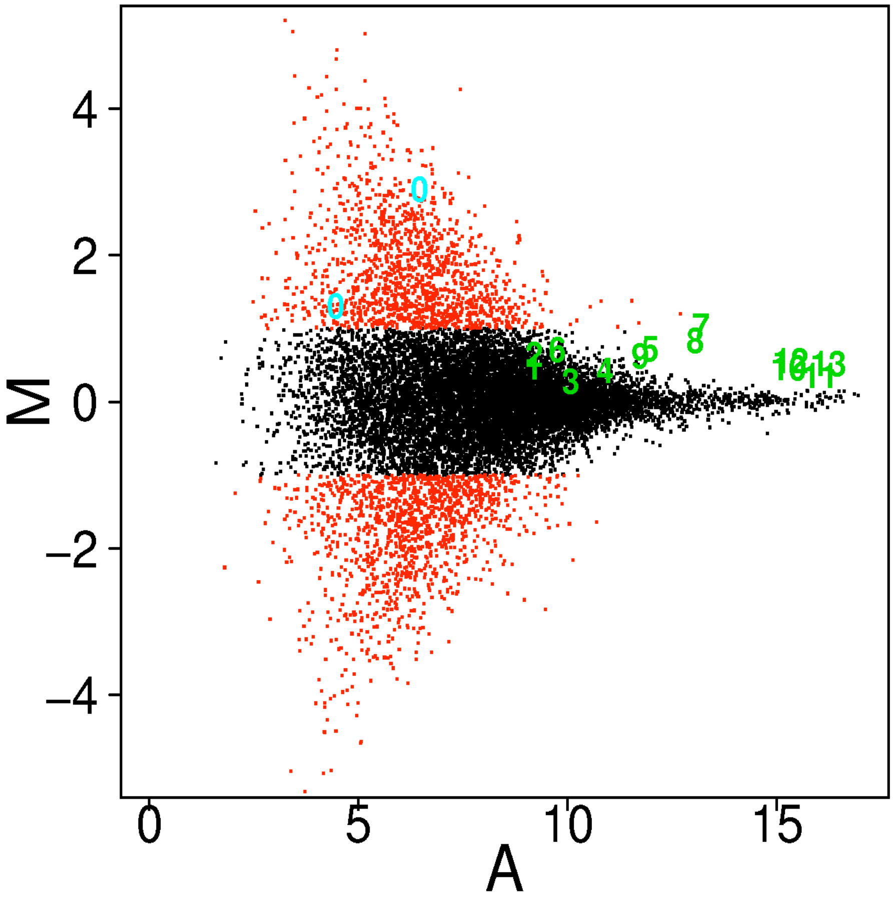
Close-up



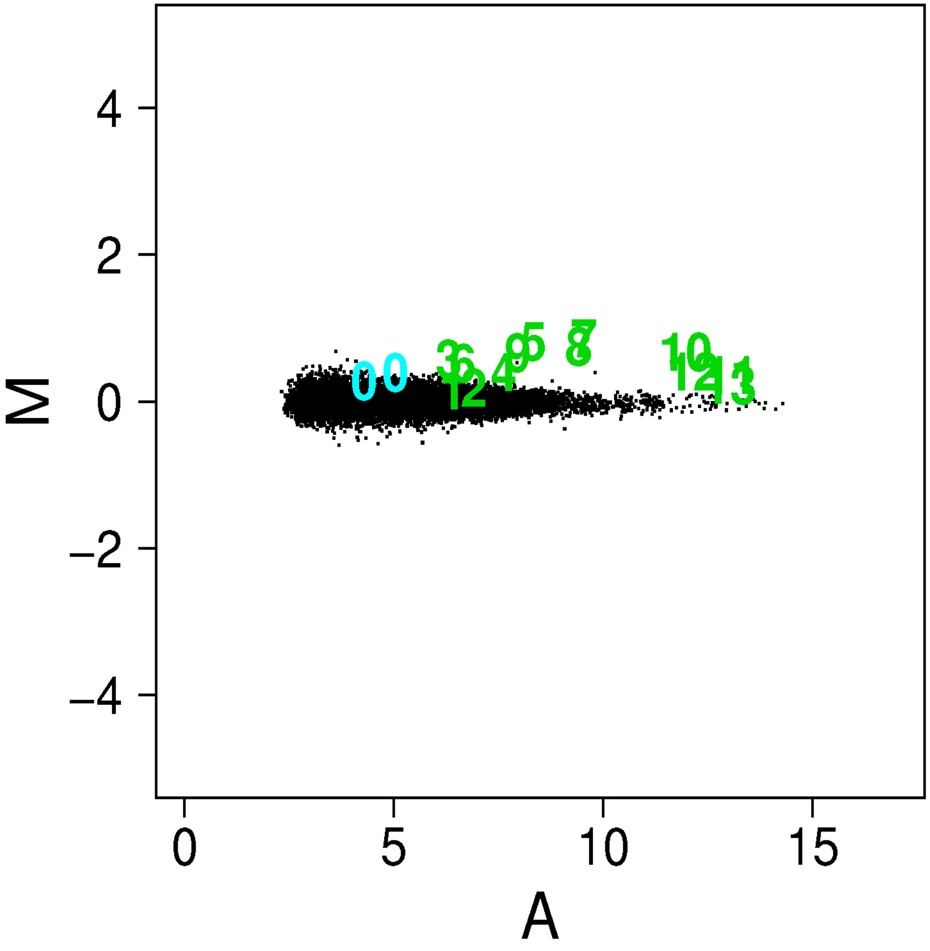
Cross-section



Does it make a difference?

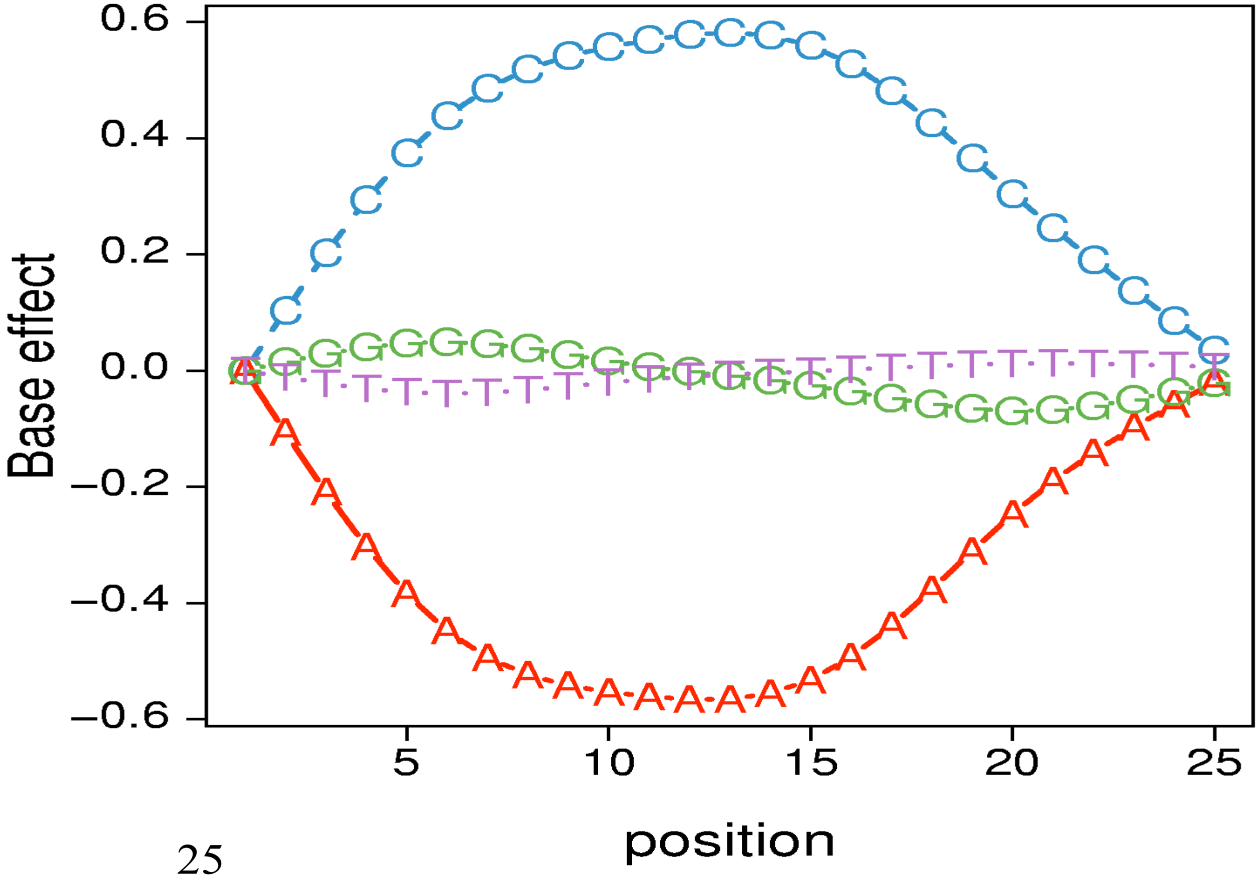


Much better precision
Slightly less accuracy



Probe Sequence

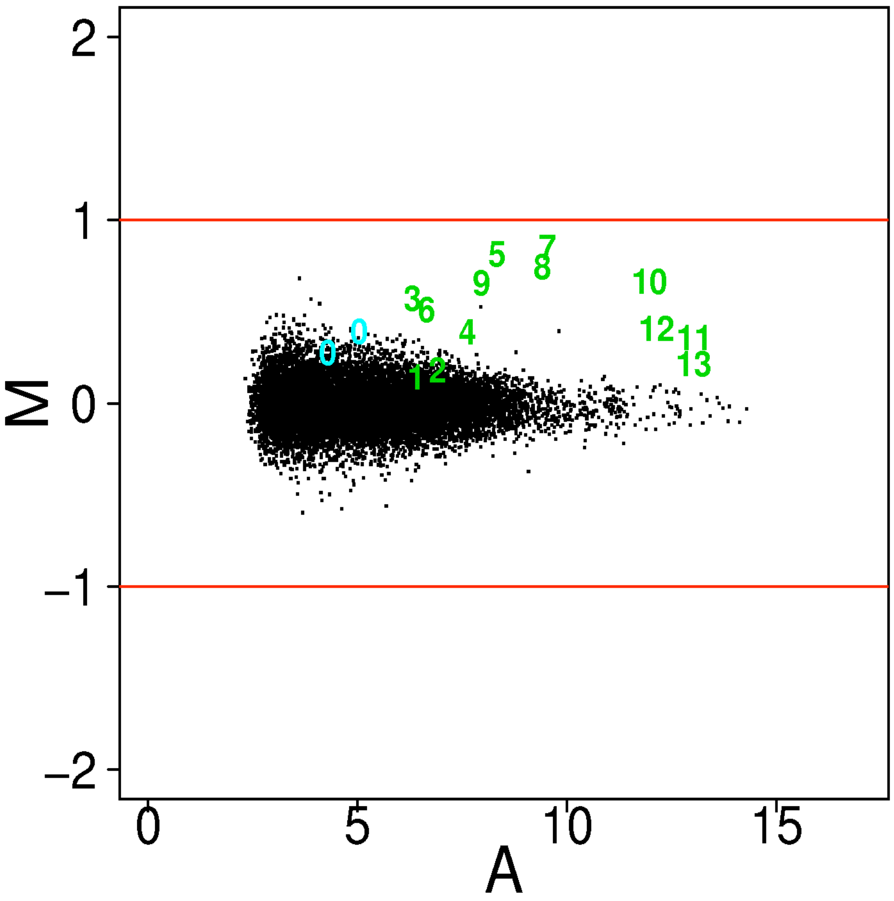
Zhang, Miles and Aldape (2003) Nature Biotech 21
Naef & Magnasco (2003) Nucleic. Acids Res. 31 7



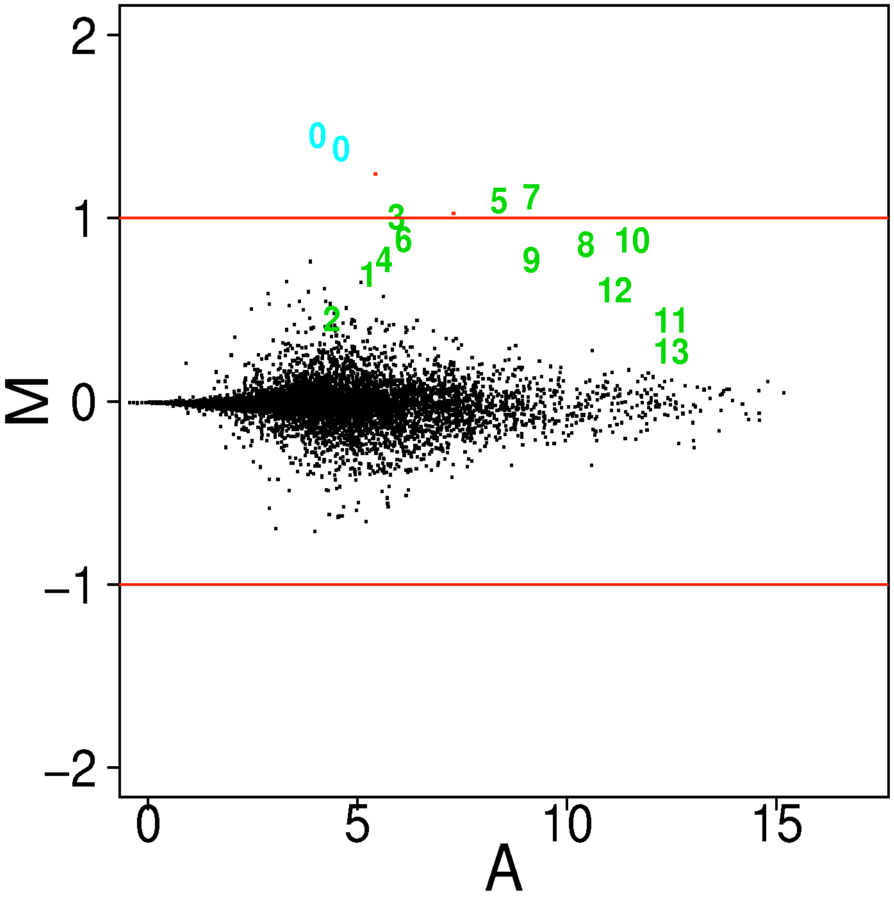
$$Affinity = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k = j}$$

$\mu_{j,k}$ ~ *smooth function of k*

Does it help?

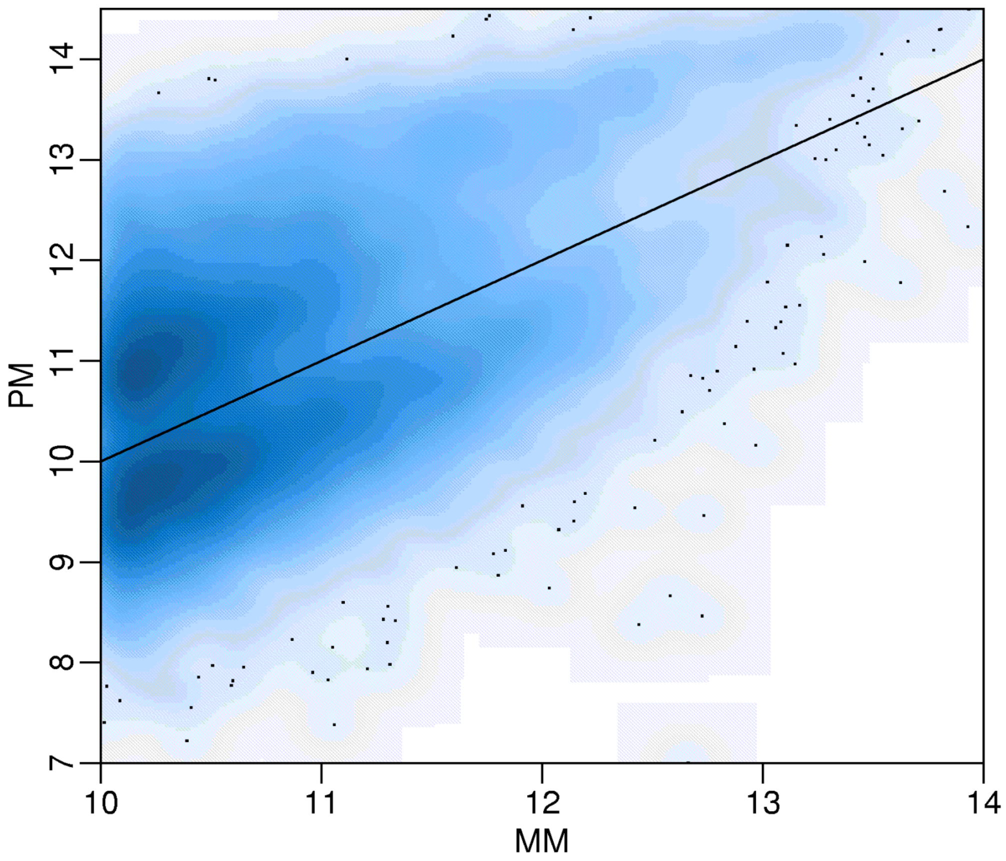


Better accuracy

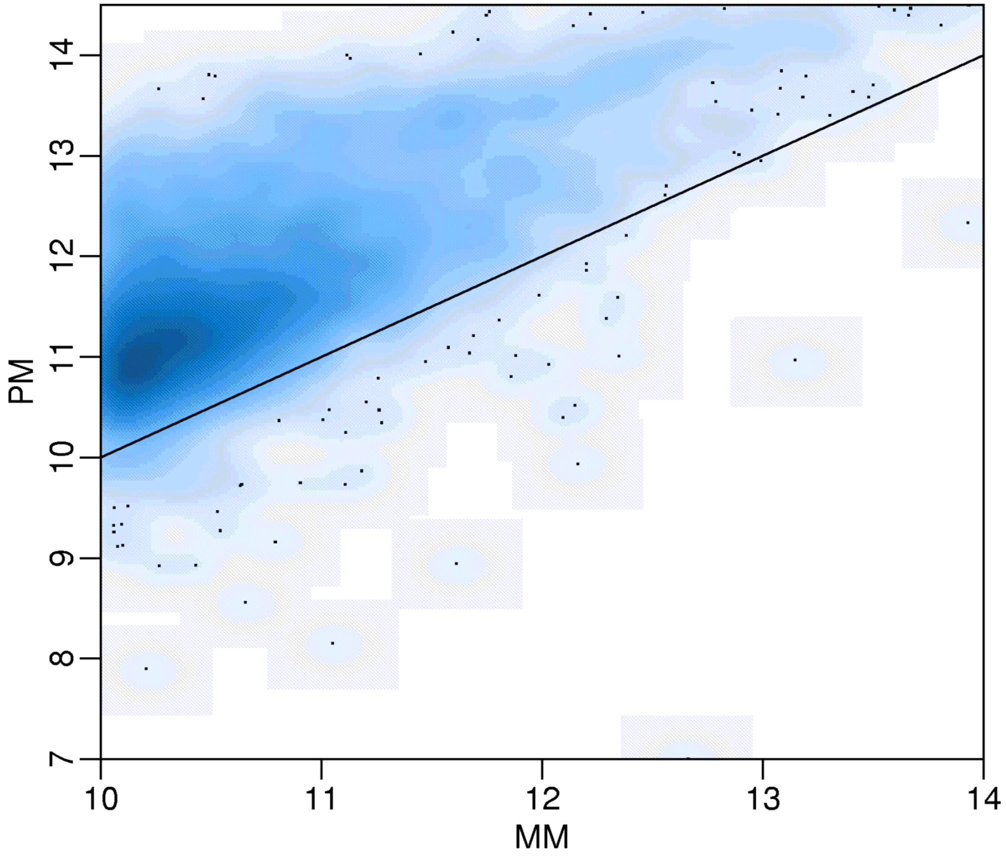


Sequence explains bimodality

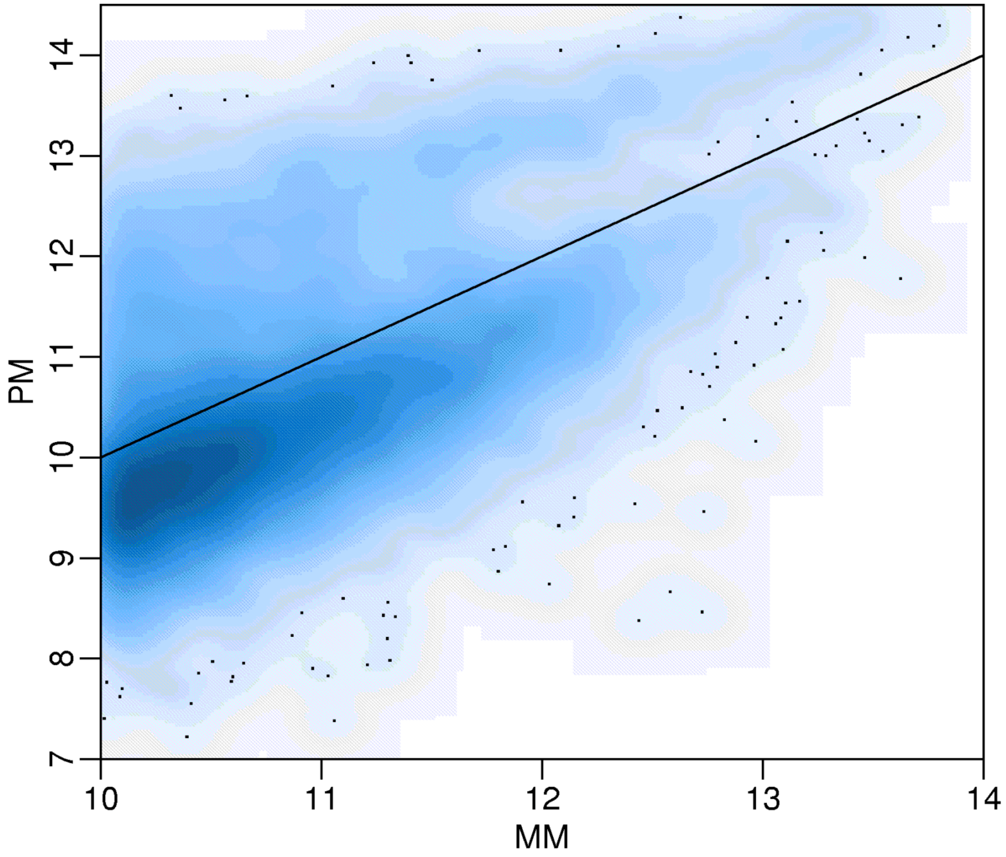
Close-up



C or *T* in the middle



A or *G* in the middle



SNP Chips

What makes some humans

hansom and others ordinary?

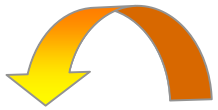
What are SNPs?

- SNPs make up 90% of all human genetic variations, and SNPs with a minor allele frequency of ≥ 1% occur every 100 to 300 bases along the human genome, on average.
- Variations in the DNA sequences of humans can affect how humans develop diseases, respond to pathogens, chemicals, drugs, etc. As a consequence SNPs are of great value to biomedical research and in developing pharmacy products.

From Wikipedia

Affymetrix SNP chip terminology

Genomic DNA



SNP

A
TAGCCATCGGTANGTACTCAATGAT
G

Perfect Match probe for Allele A **ATCGGTAGCCATTCATGAGTTACTA**

Perfect Match probe for Allele B **ATCGGTAGCCATCCATGAGTTACTA**

Genotyping: answering the question about the two copies of the chromosome on which the SNP is located:

Is a person **AA** , **AG** or **GG** at this Single Nucleotide Polymorphism?

In summary: probe level data

- Two alleles
- Two directions
- Two types (PM,MM)
- Up to 7 locations of the SNP in the probe

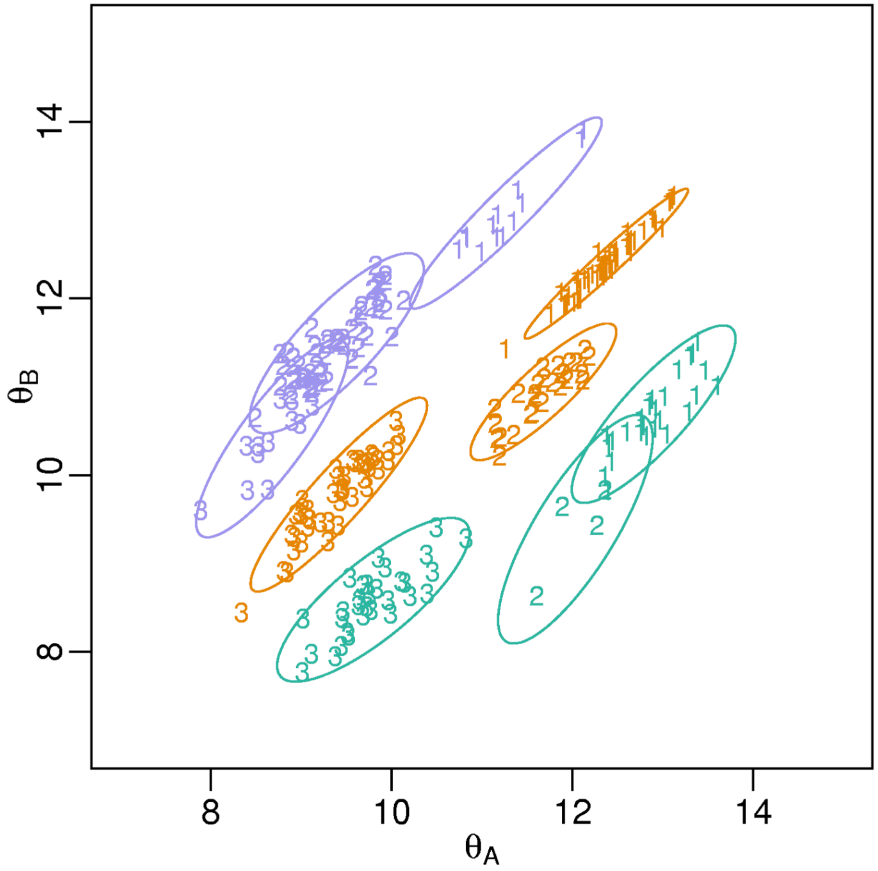
Notation

- Once we are done with first part of preprocessing we have the following:

  θ_A and θ_B proportional to log of the amount of fragments from allele A and B respectively

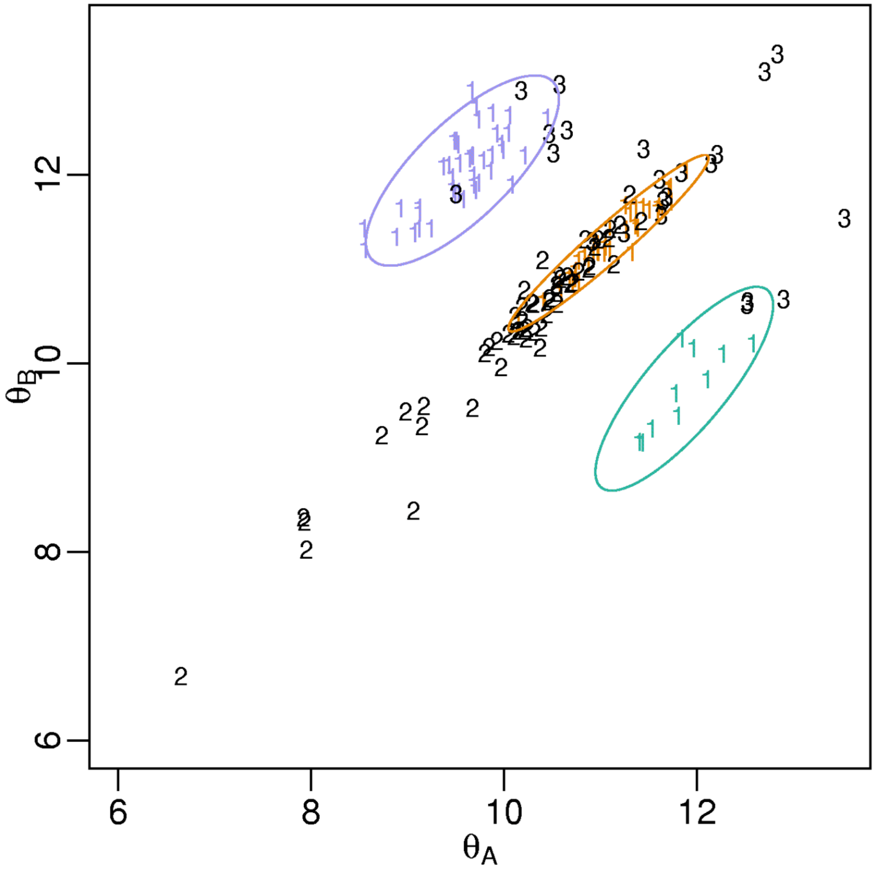
In principal these can only be (log of) 0, x, or 2x, but we know better than to believe this.. In fact we know not to expect the same cut-off to work for all SNPs

It’s not easy



This picture shows that most the information is in the left right diagonal direction, i.e. in the log-ratios

Lab Effect



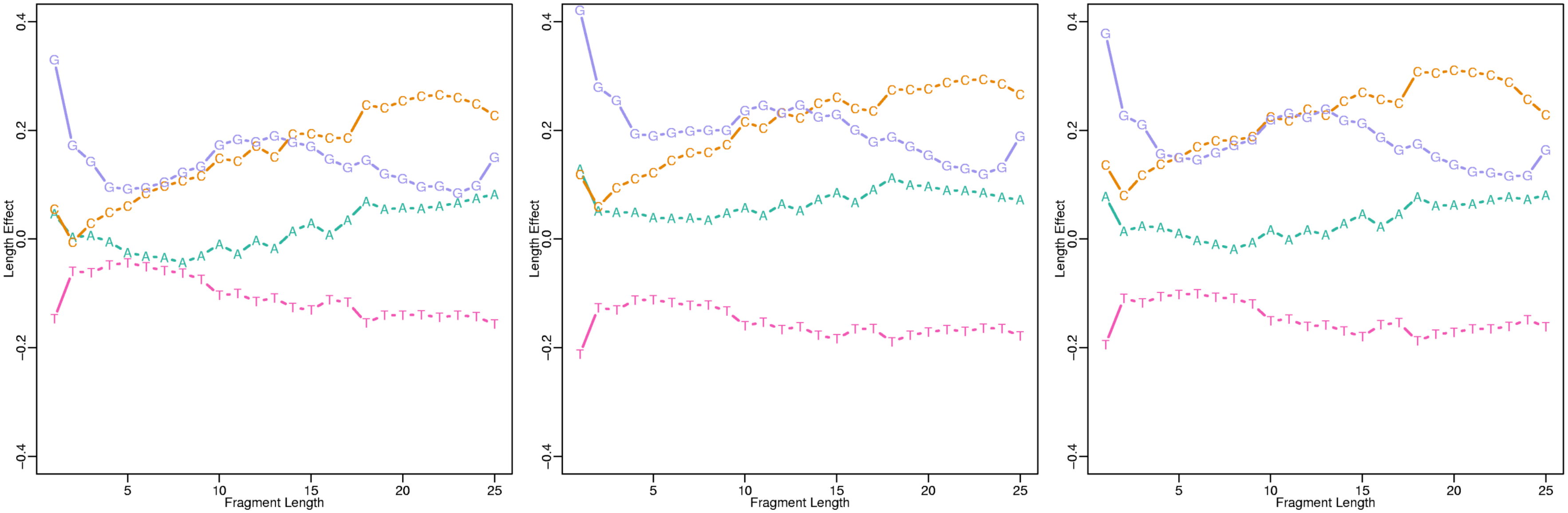
Why is this?

- Our guess is that the PCR step introduces a lot of SNP to SNP variation

- We have proxies for measuring PCR effect: fragment sequence and fragment length

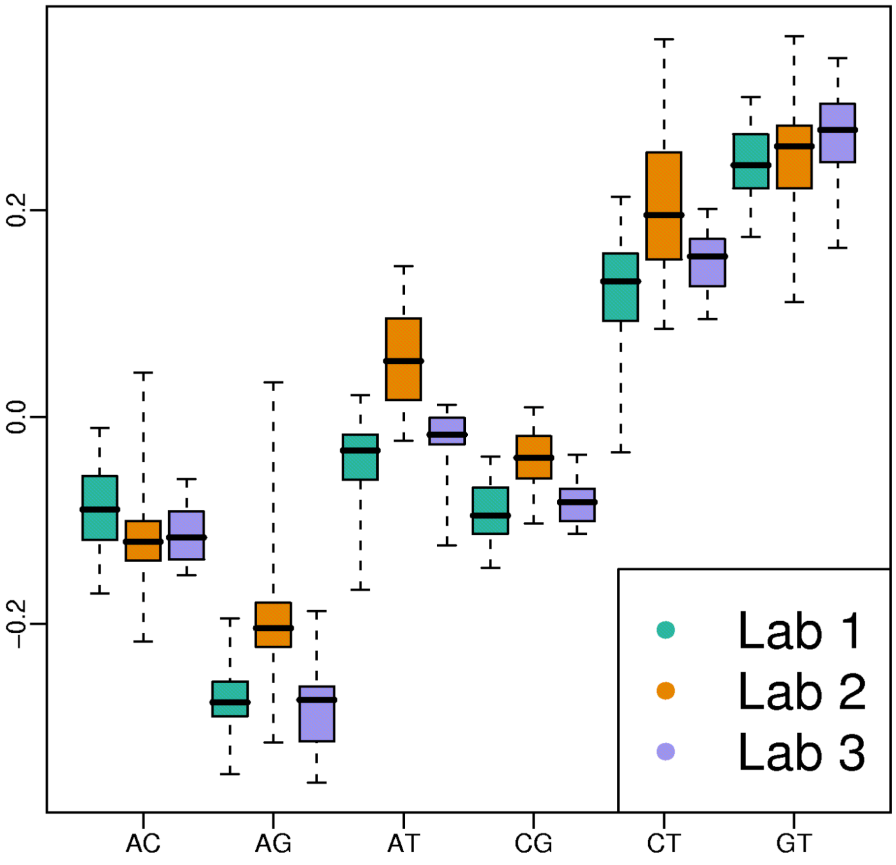
- We can examine the fragment sequence via the probe sequence

Sequence effect

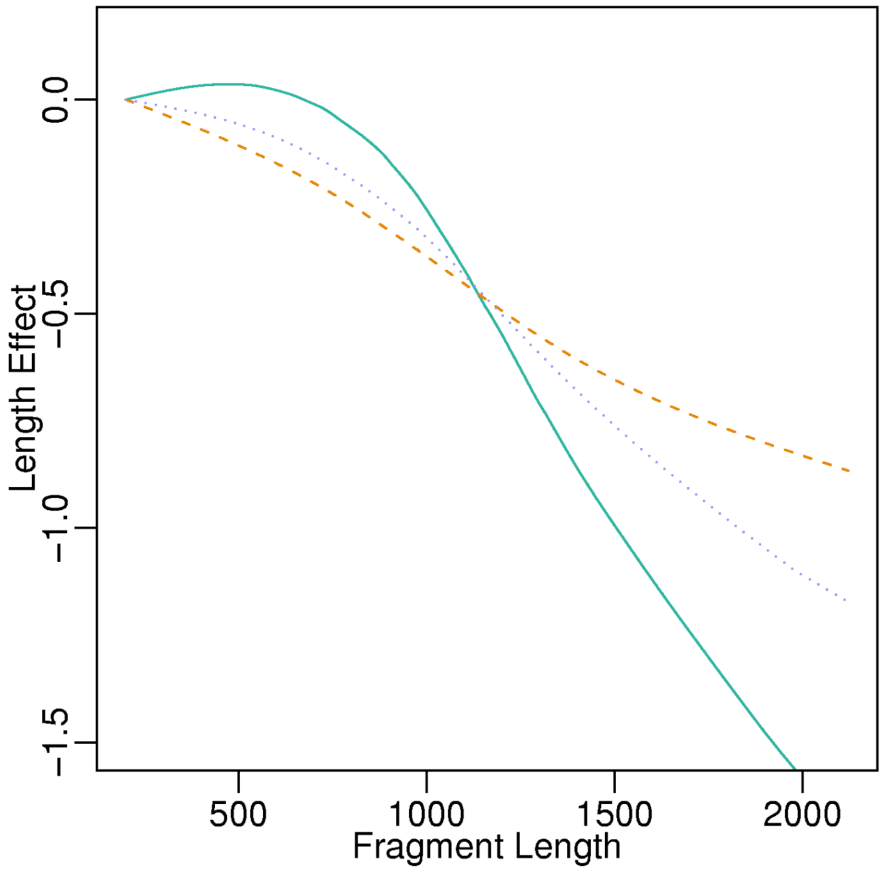


Sequence Effect ctd

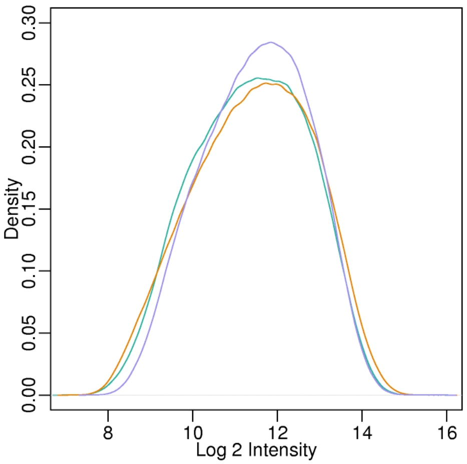
M



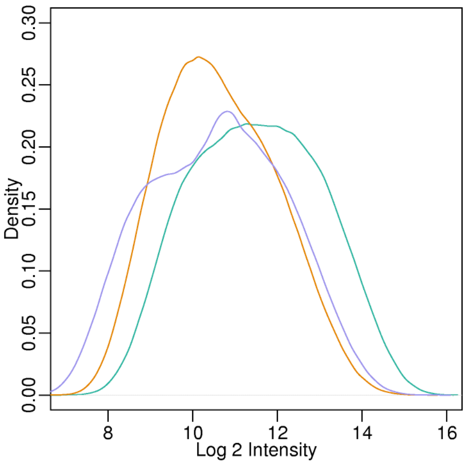
Different Labs



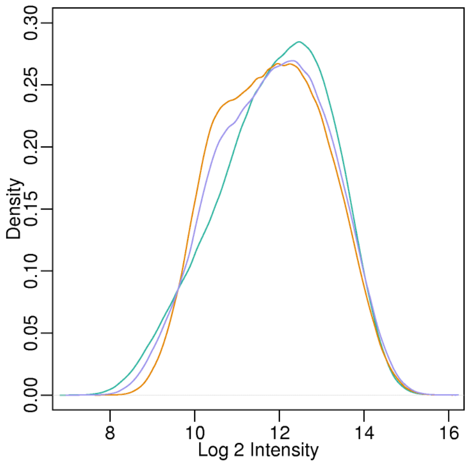
Need for Norm



Lab 1



Lab 2

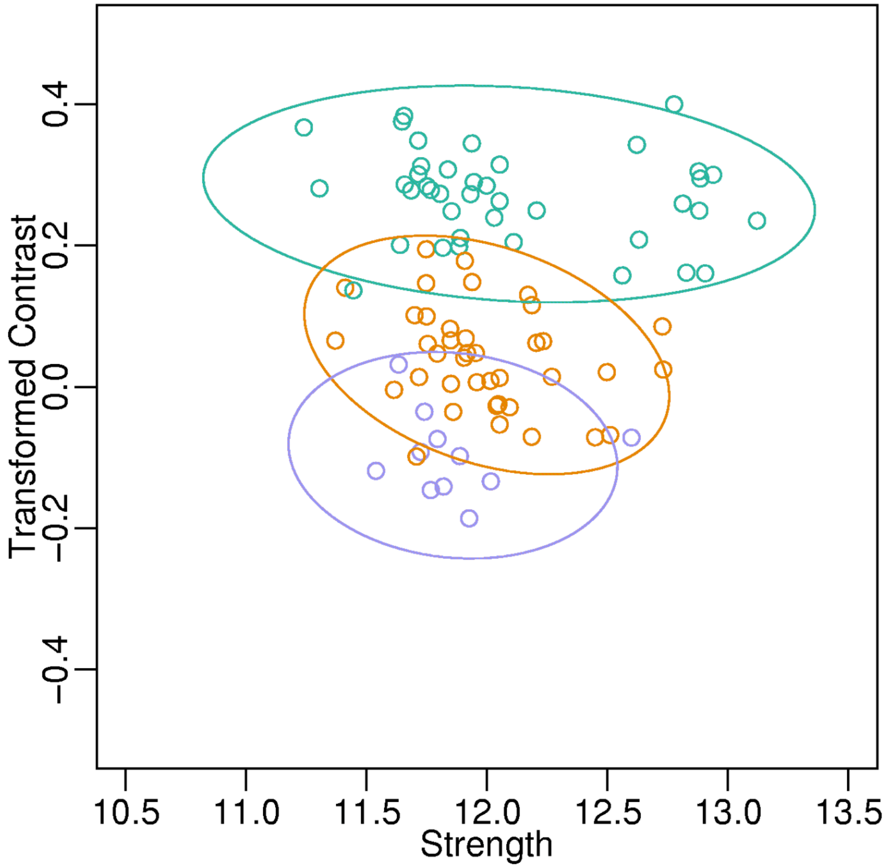


Lab 3

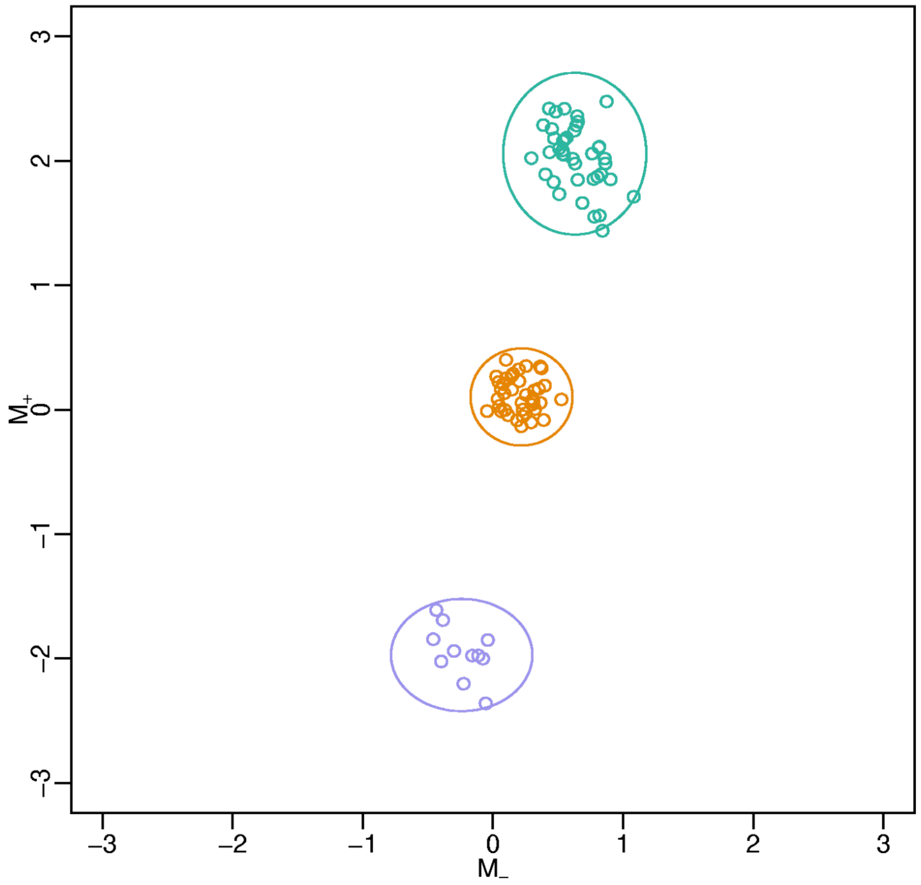
Normalization

- We normalize/summarize using RMA (no BG correction) after correcting for sequence and length effects on the log intensities
- We then examine log-ratios
- We keep sense and antisense separate

BRLMM for a particular SNP

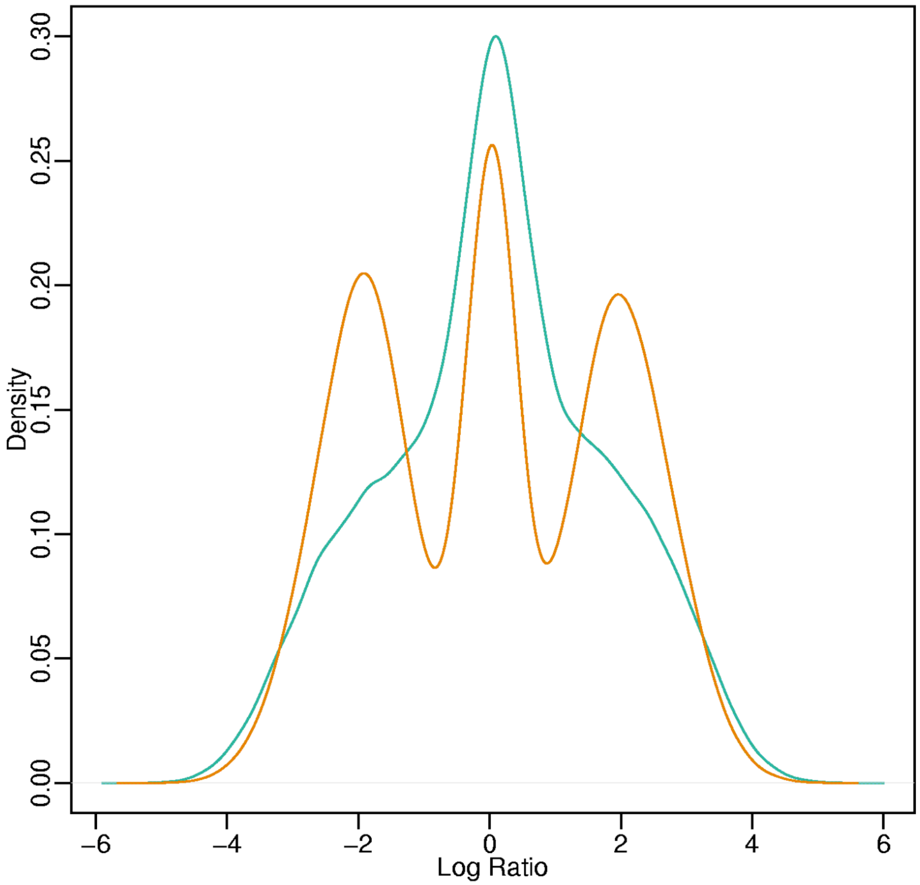


Temporarily disabled probes?

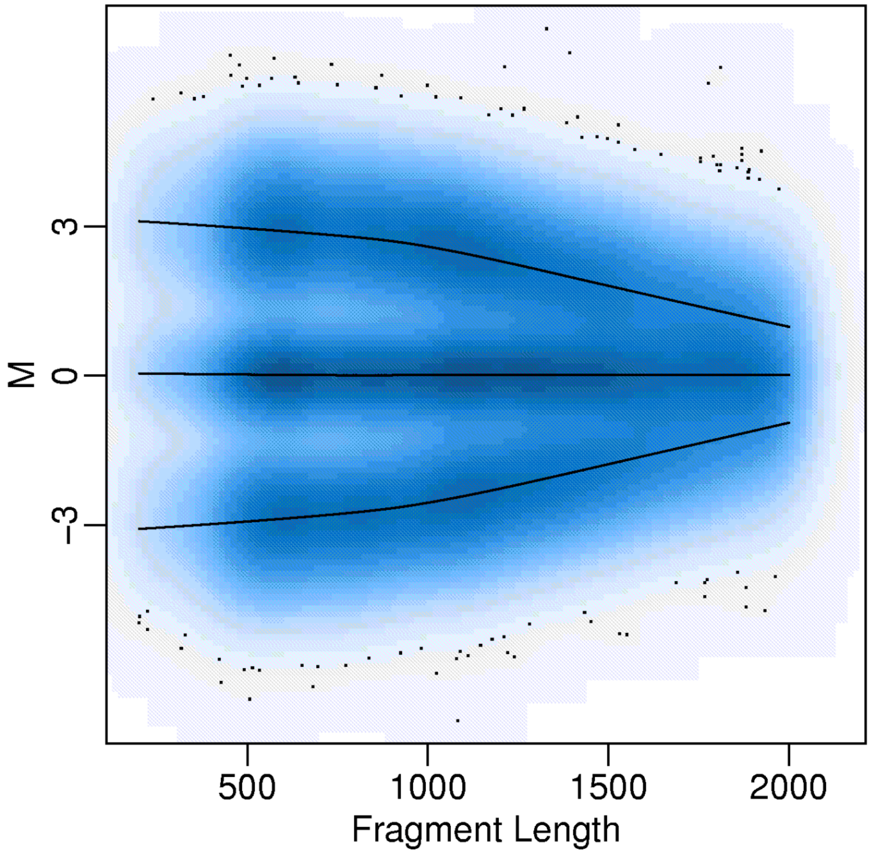


Log-ratio biases persist

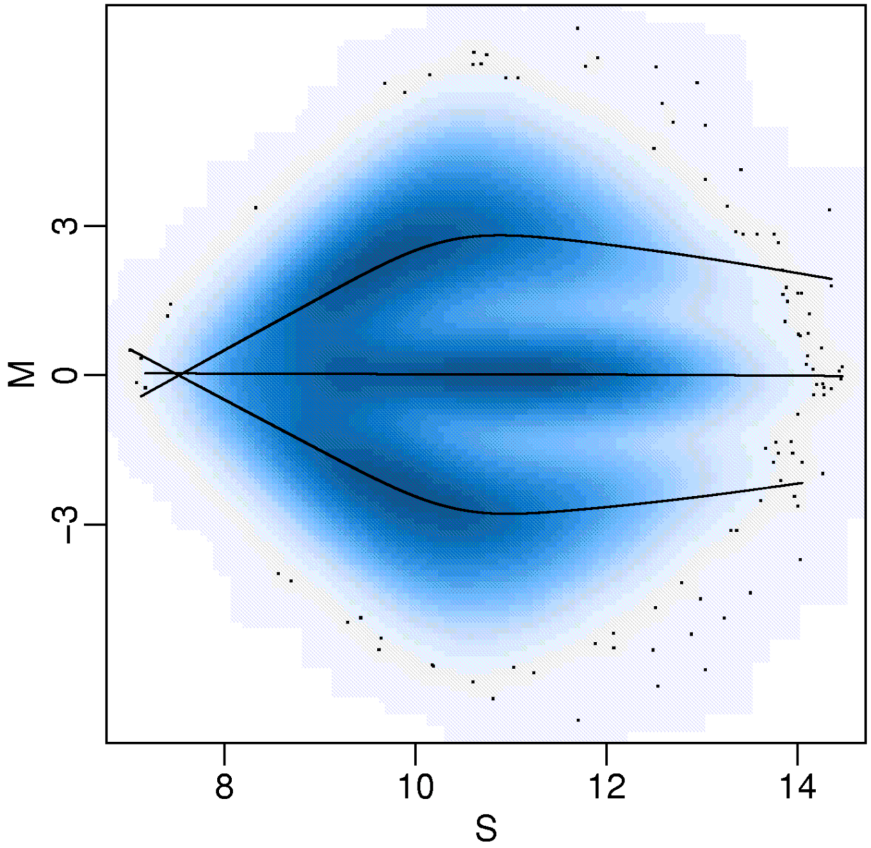
Different arrays, different cut-offs



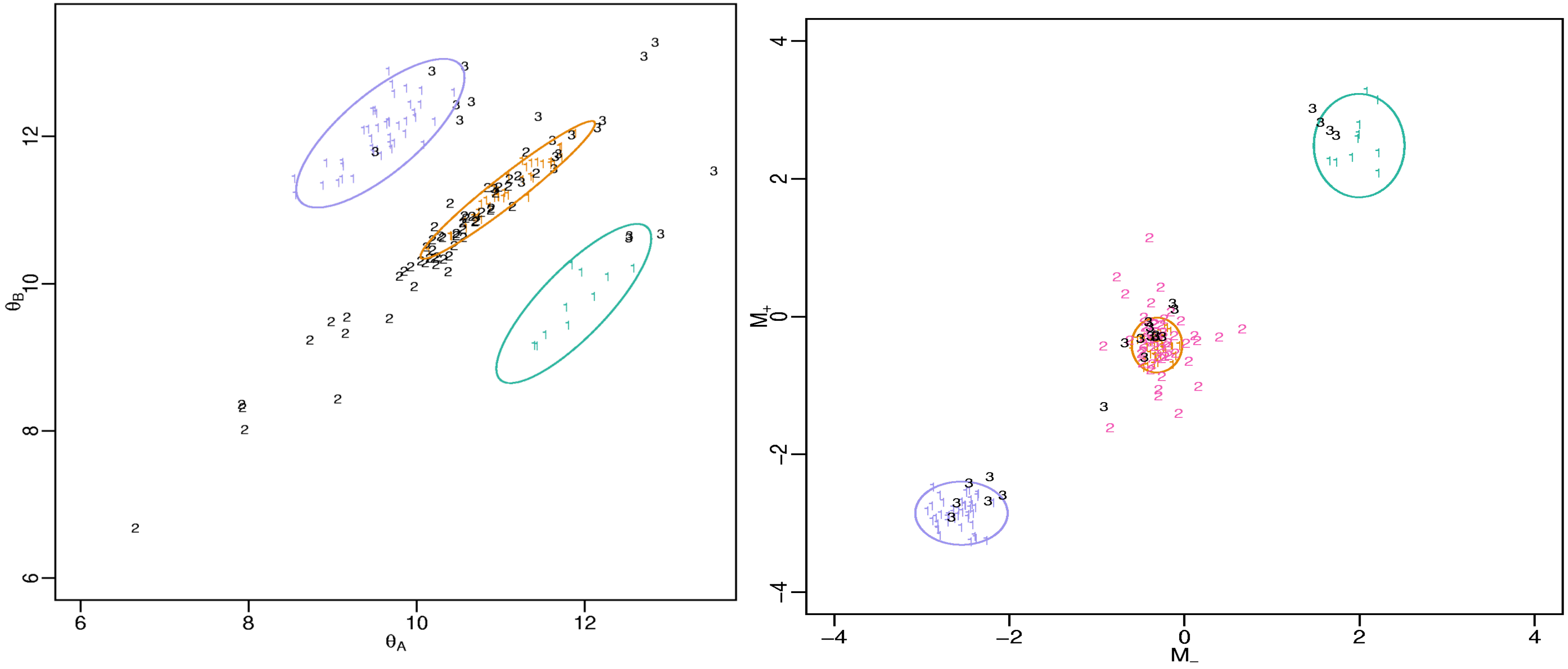
Length effect on M



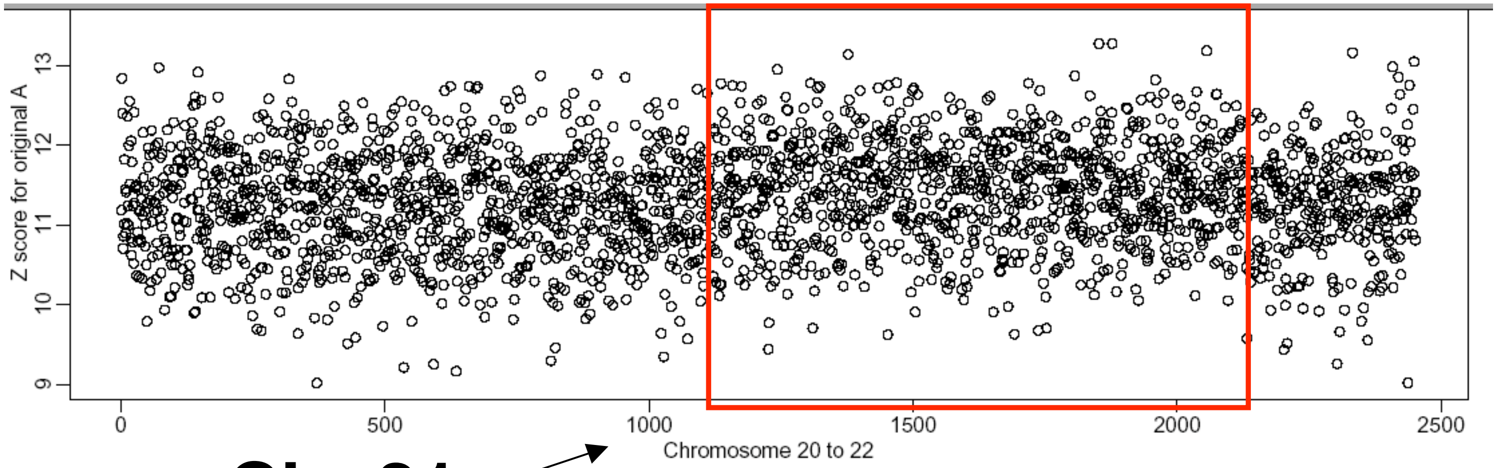
Intensity effect on M



After our normalization

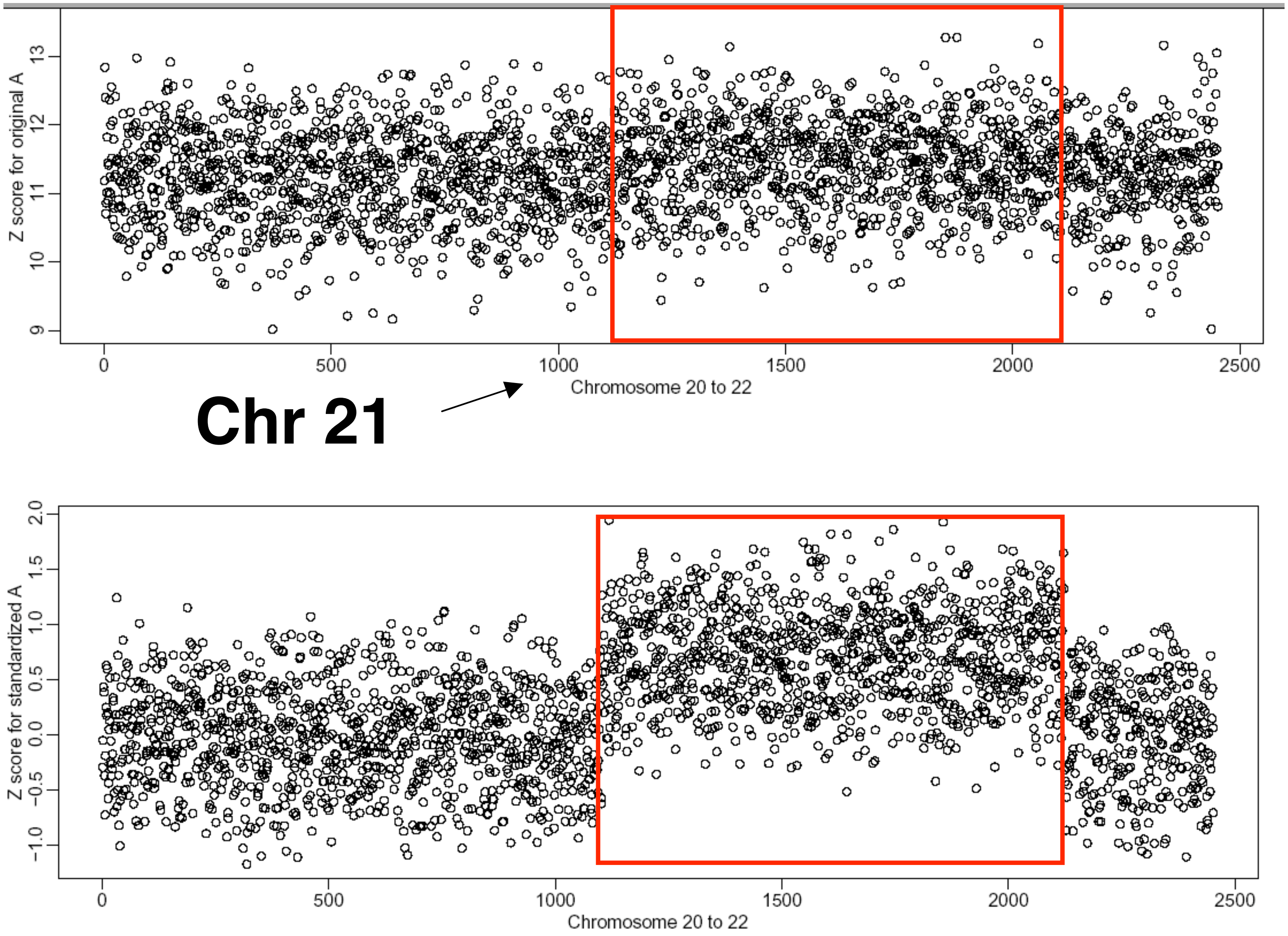


Don’t forget copy number



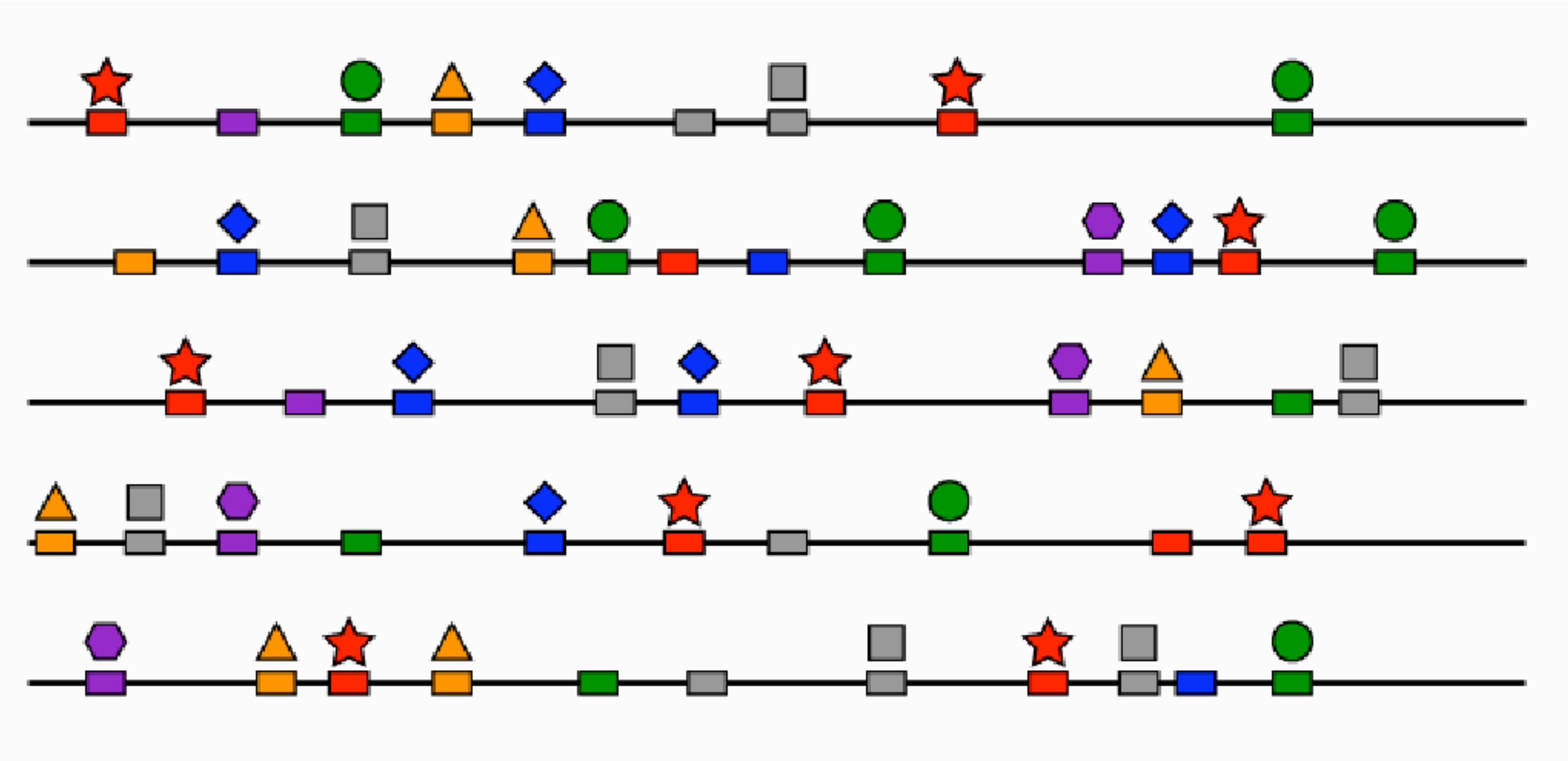
Chr 21

Don’t forget copy number

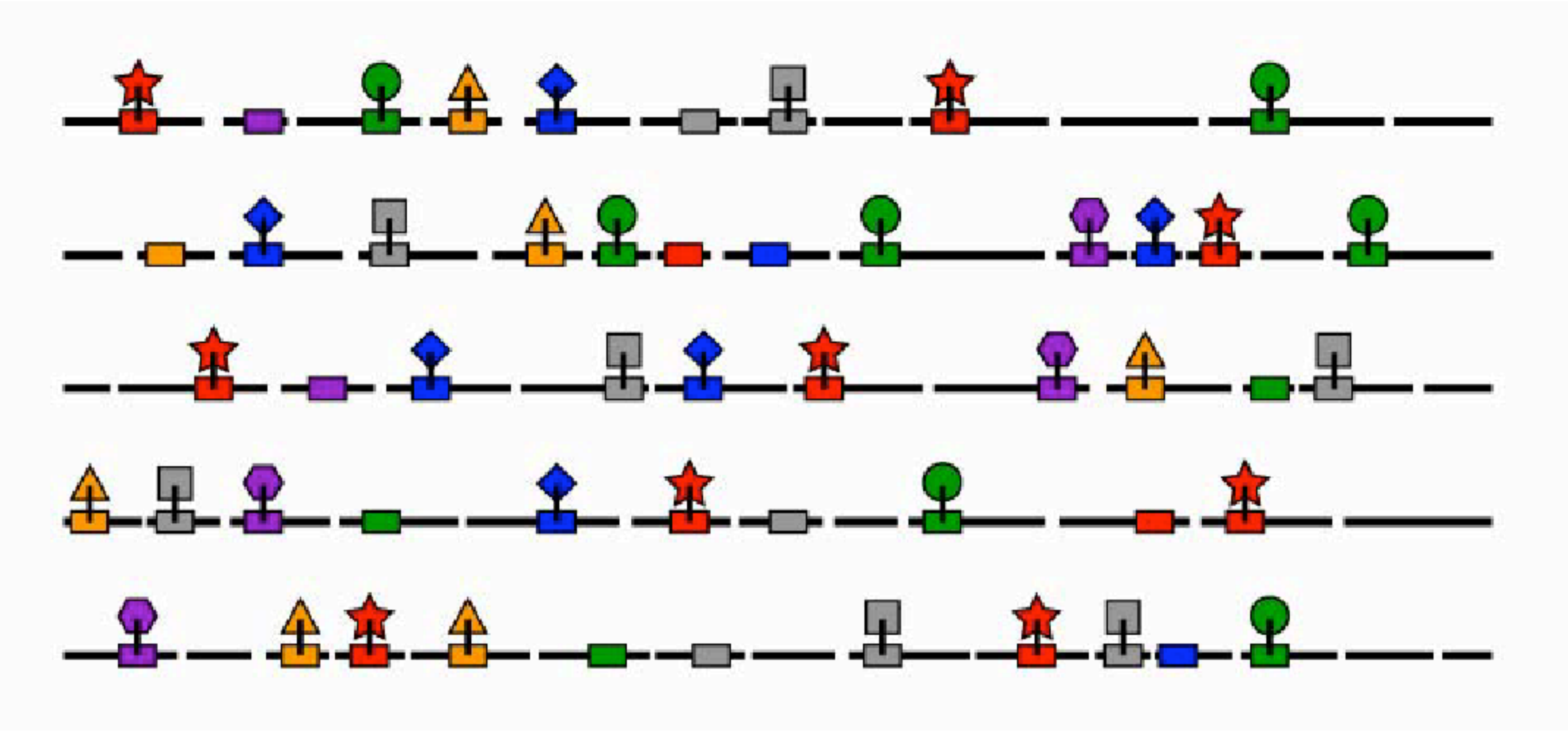


Tiling arrays

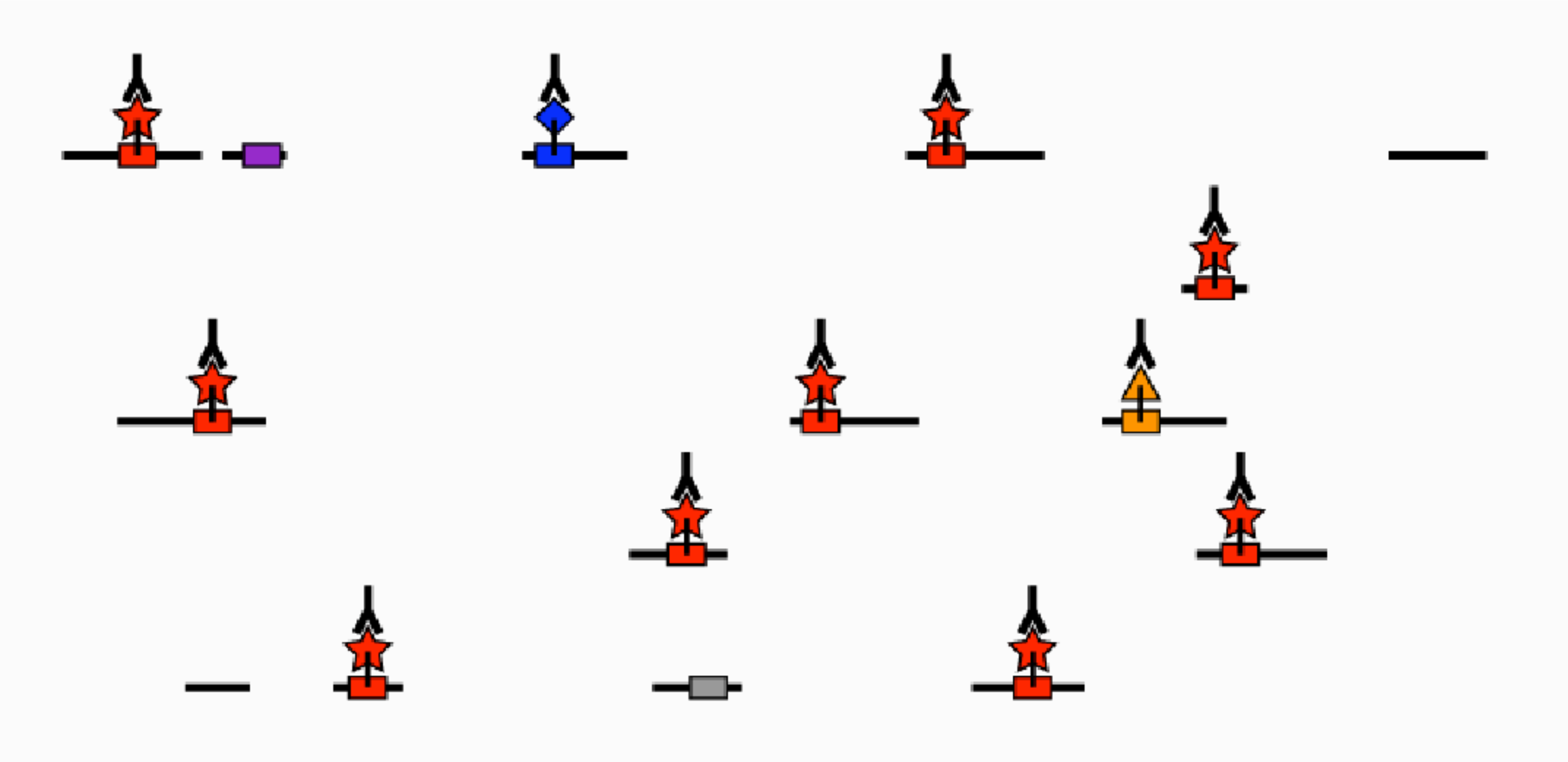
Genome



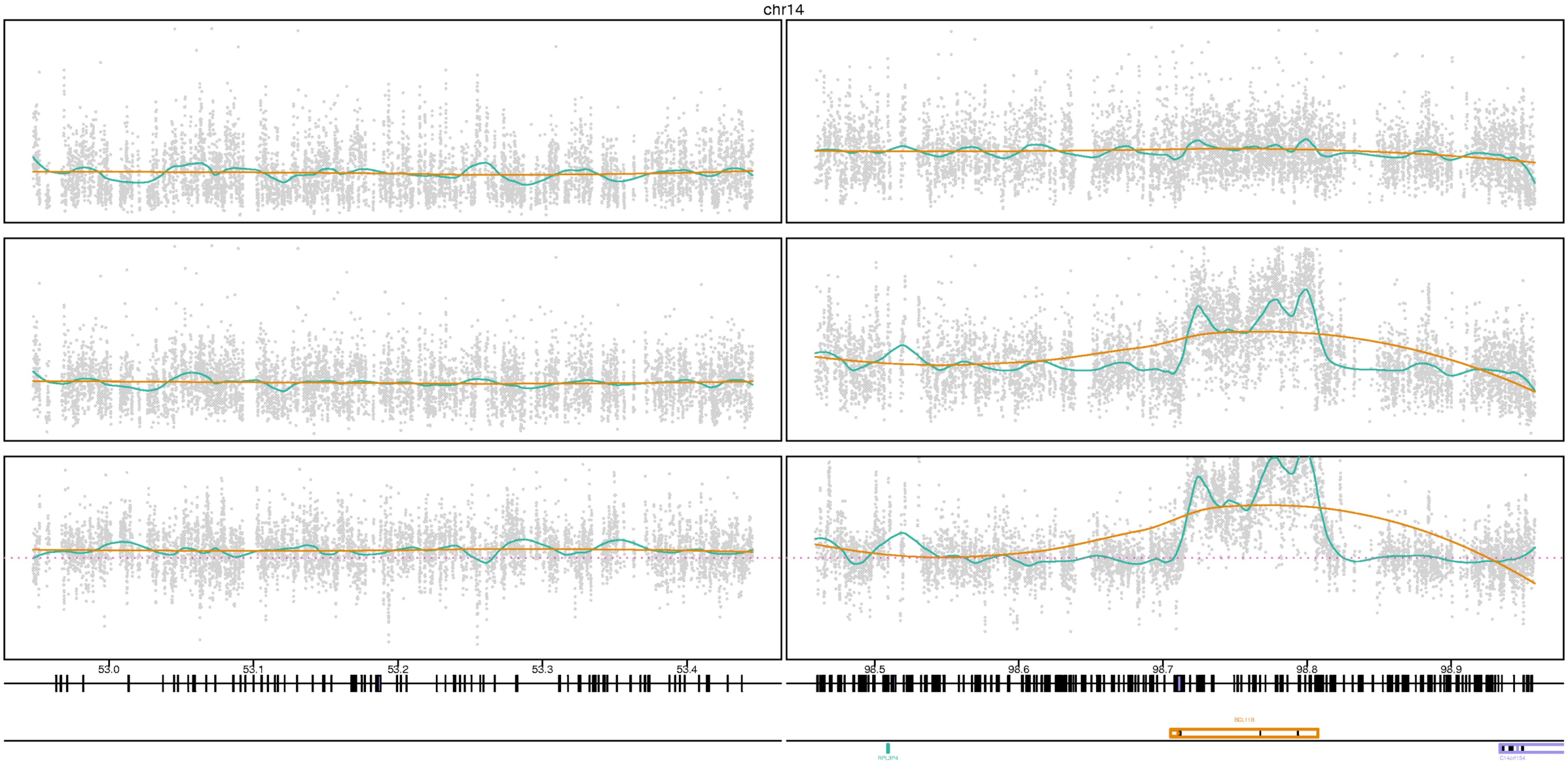
Sonification



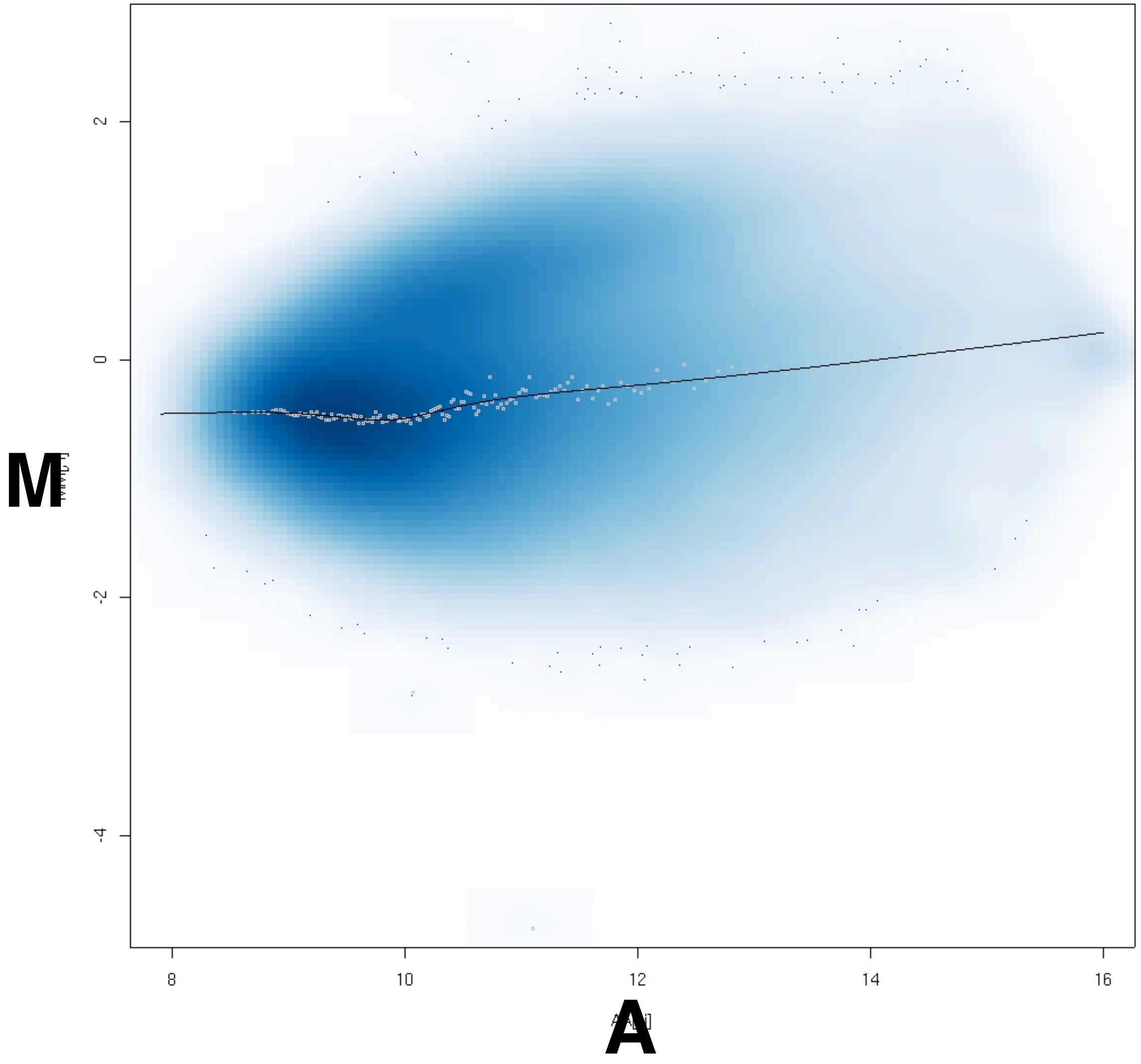
Filtering



Looking for bumps



Normalization is harder

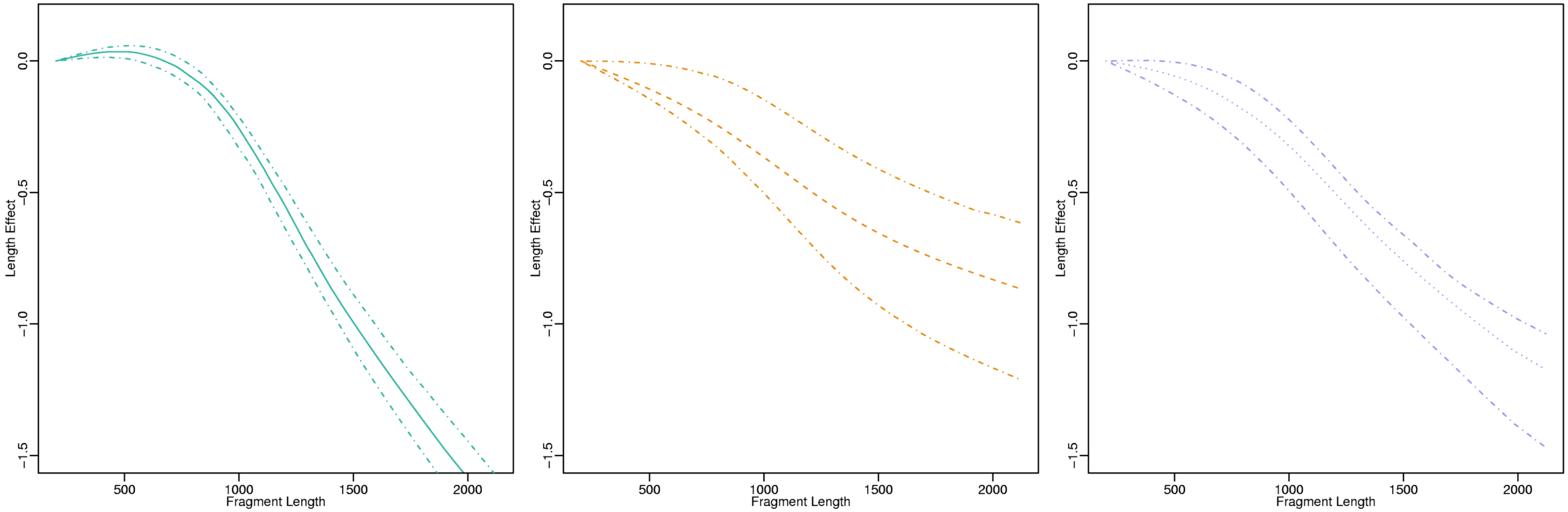


Conclusions

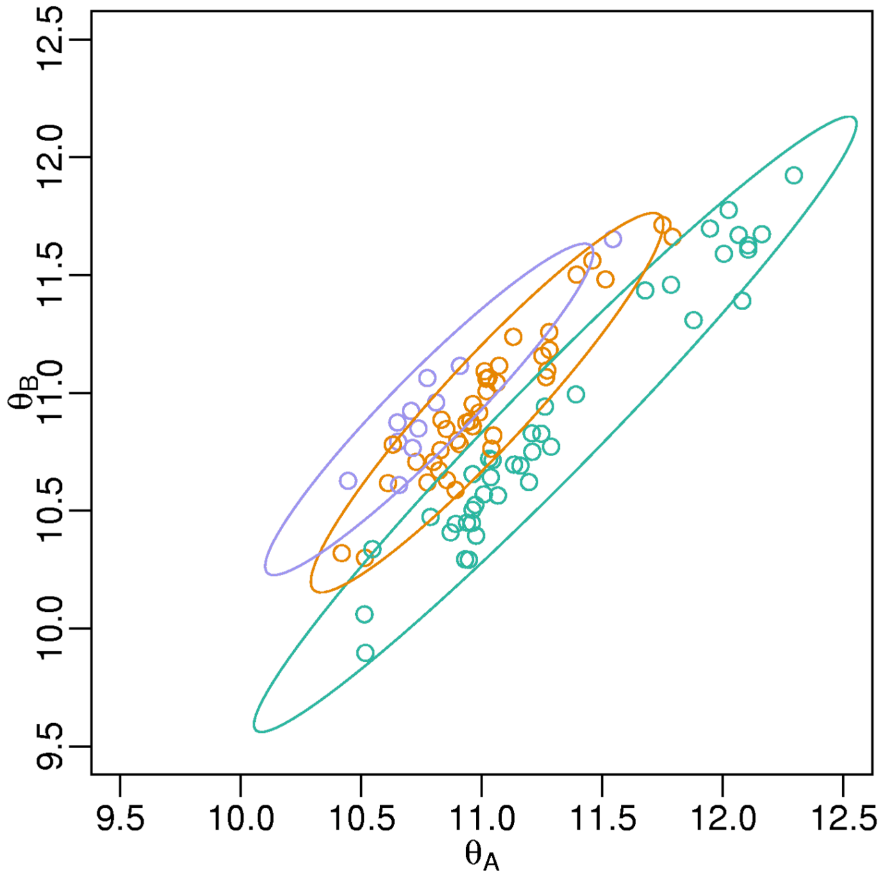
- Preprocessing algorithms make implicit assumptions that can greatly affect bottom line results
- Important to understand background noise and probe-effects to understand how/why this happens
- Better understanding can improve detection limits

Supplemental Slides

Fragment length effect



“Broken” probes (RLMM)

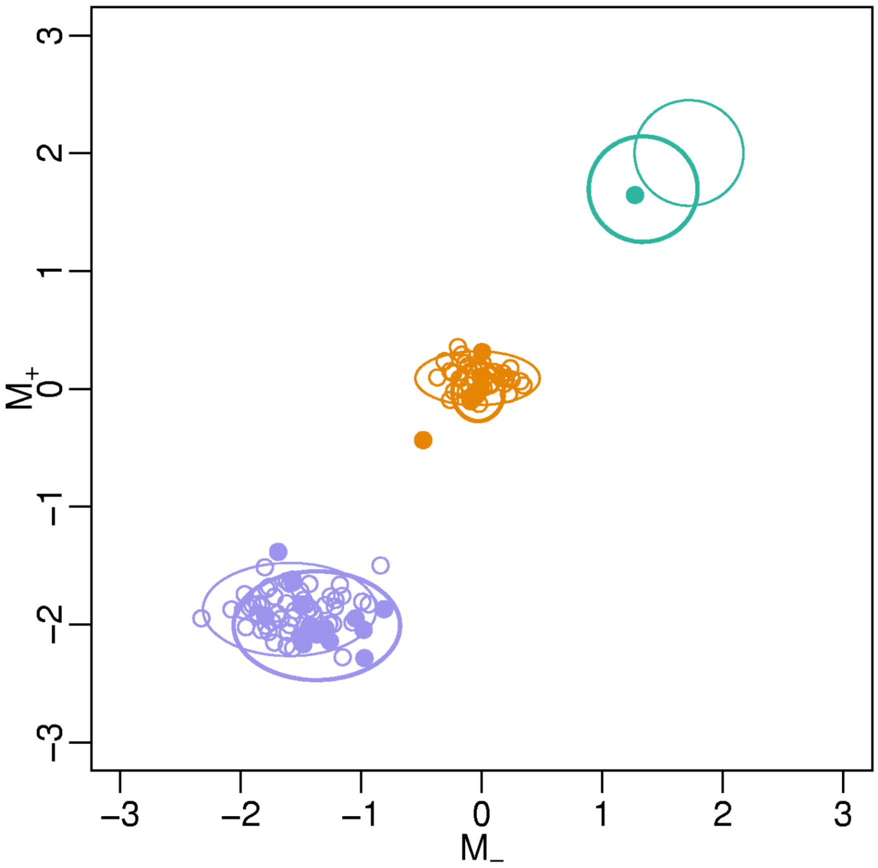


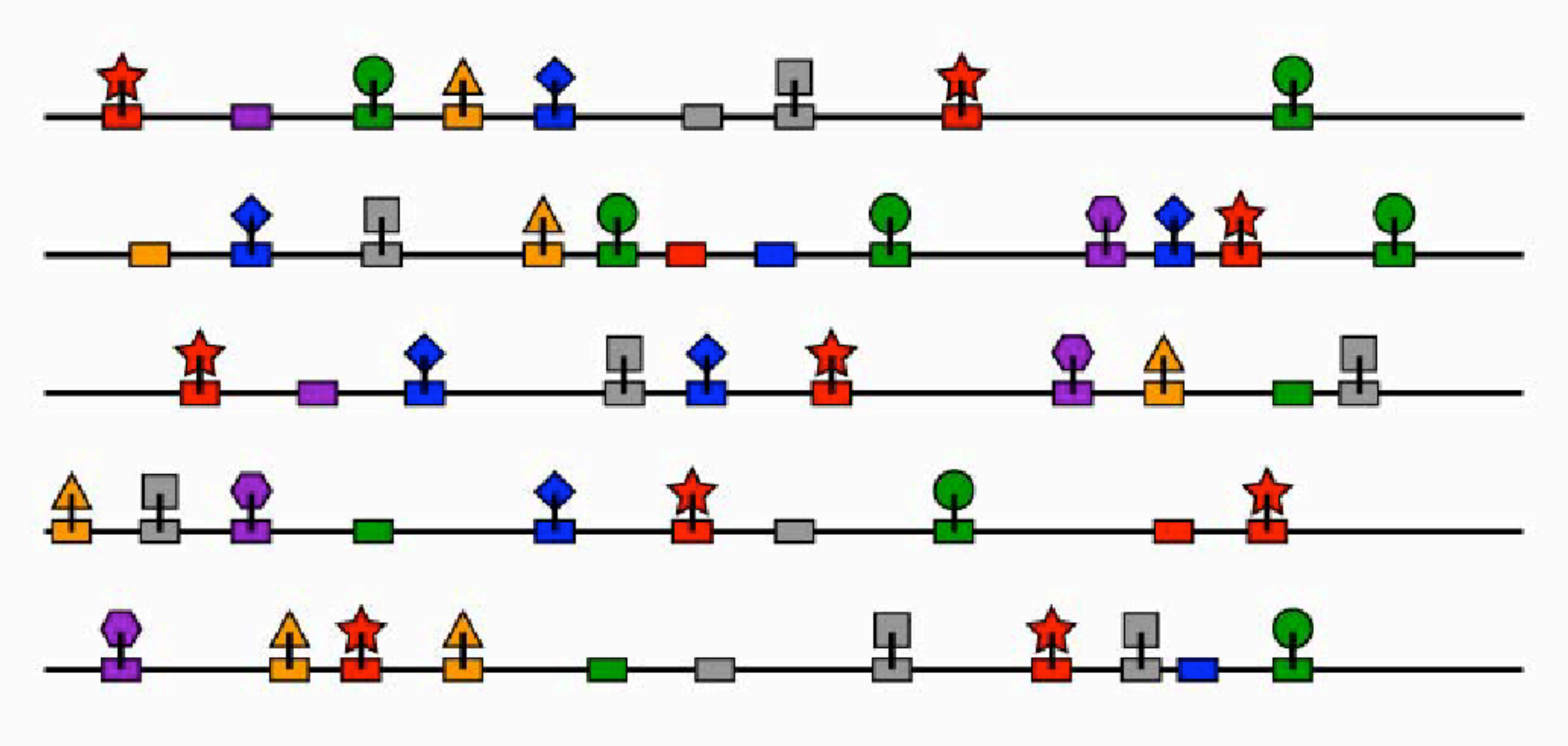
Preprocessing model motivates genotype algorithm

$$[M_{i,j,s}|Z_{i,j}=k,m_{i,k,s}]=f_{j,k}(X_{i,j,s})+m_{i,k,s}+\varepsilon_{i,j,k,s}.$$

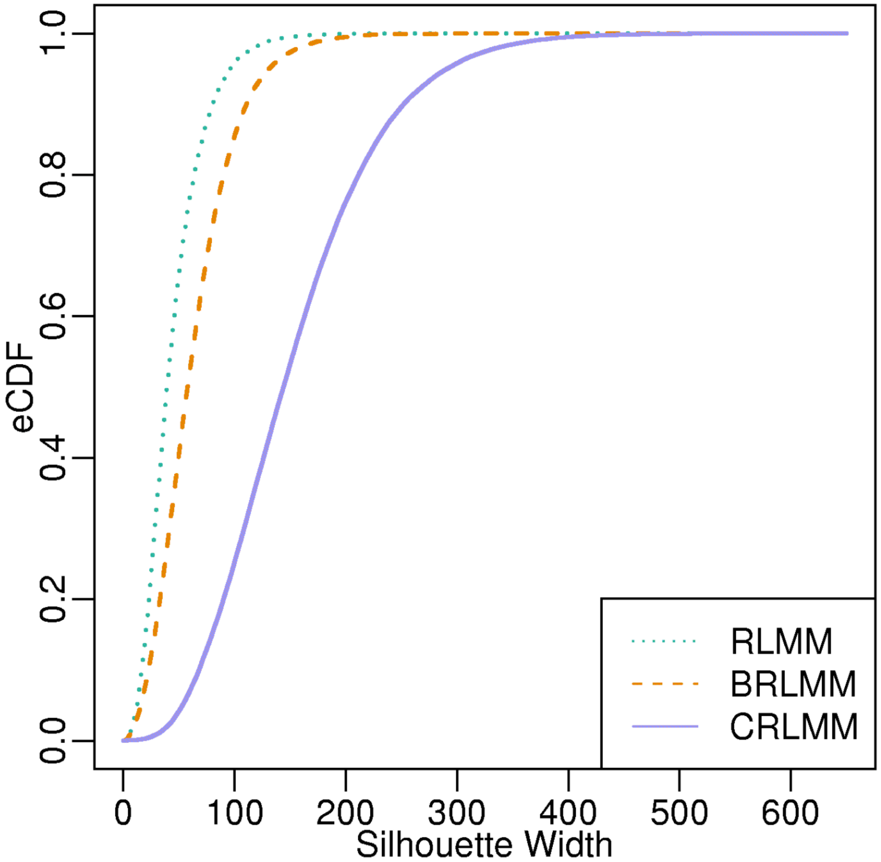
- Array denoted with j
- Shift in cluster center denoted with m
- We assume m is normal
- Use training data to estimate m
- Use empirical bayes approach for cases with few data points

Example

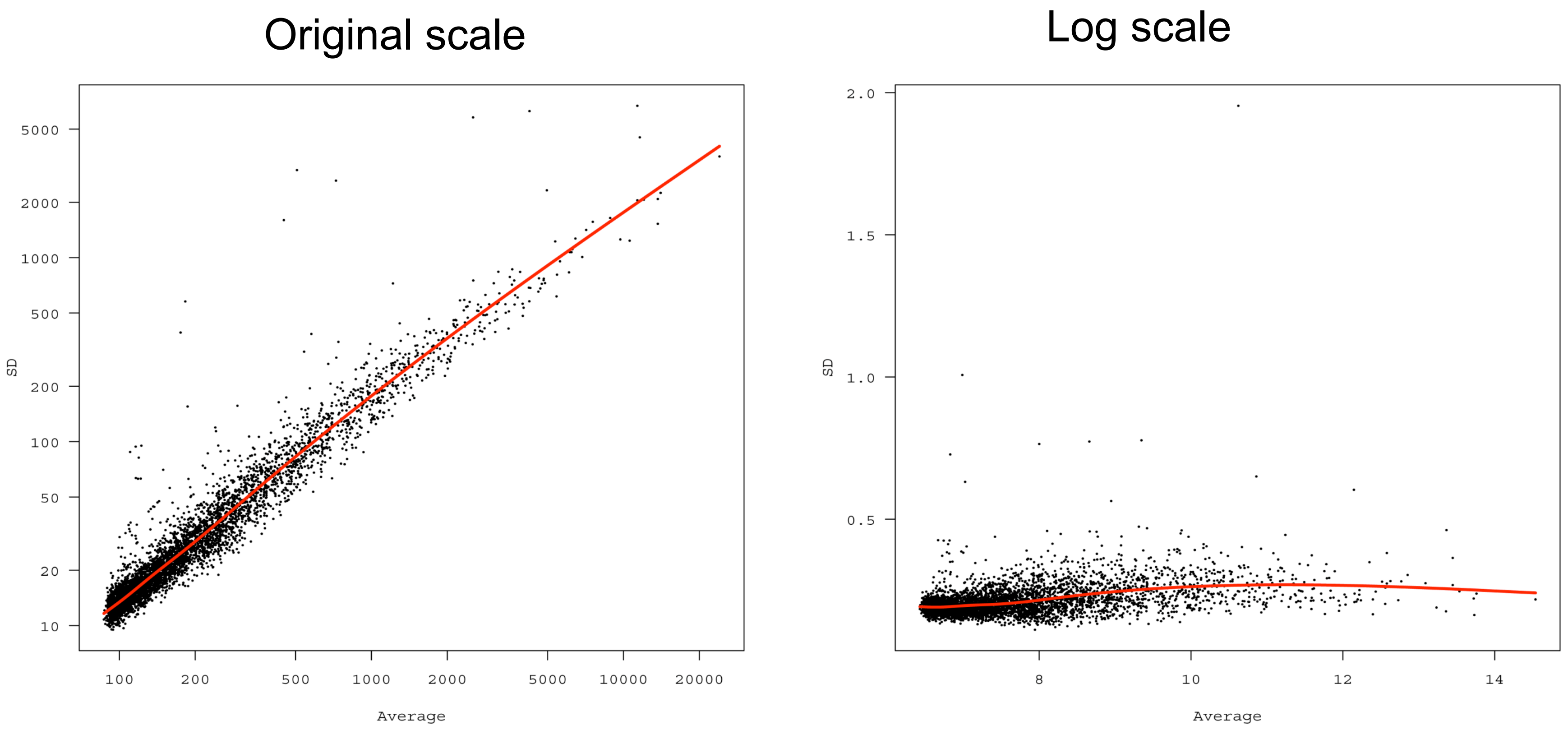




General Improved Separation



Why logs?



Original scale

Log scale

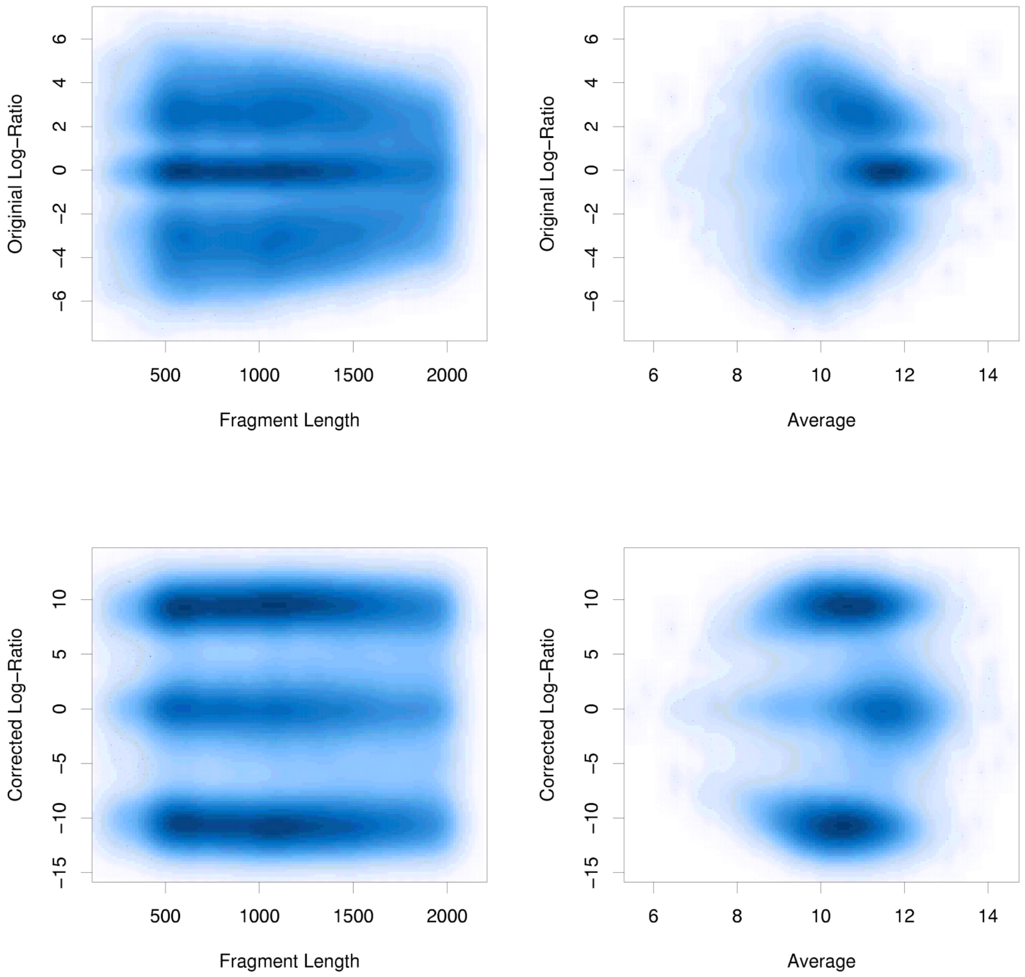
SD versus Avg plots

Use mixture model to fix this

$$[M_i|Z_i = k] = f_k(X_i) + \varepsilon_{i,k}$$

- SNP denoted with I
- Z is true, so k = AA, AB or BB
- X are covariates that cause bias

After fix


Originial Log−Ratio
6
4
2
0
−2
−4
−6
500
1000
1500
2000
Fragment Length
Original Log−Ratio
6
4
2
0
−2
−4
−6
6
8
10
12
14
Average
Corrected Log−Ratio
10
5
0
−5
−10
−15
500
1000
1500
2000
Fragment Length
Corrected Log−Ratio
10
5
0
−5
−10
−15
6
8
10
12
14
Average

Tiling strategy

SNP 0 position

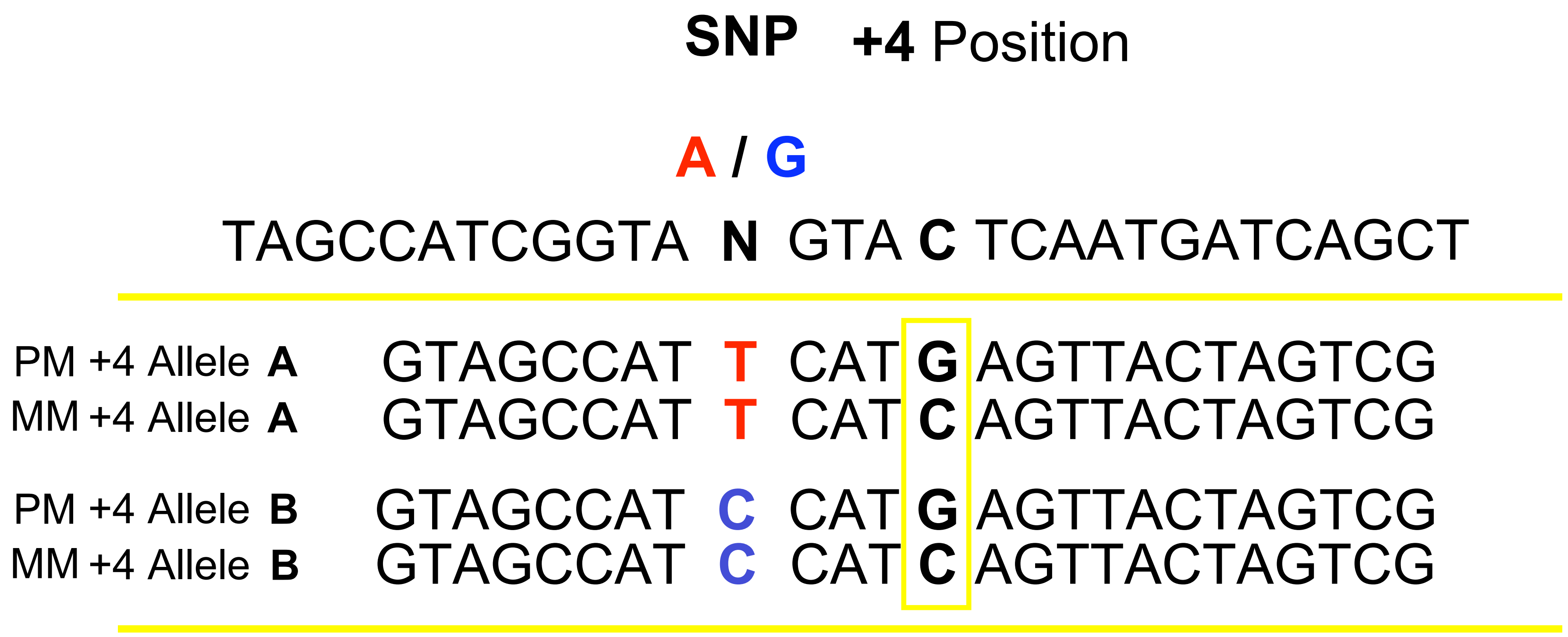
A / G

TAGCCATCGGTA **N** GTACTCAATGAT

PM 0 Allele **A**	ATCGGTAGCCAT	**T**	CATGAGTTACTA
MM 0 Allele **A**	ATCGGTAGCCAT	**A**	CATGAGTTACTA
PM 0 Allele **B**	ATCGGTAGCCAT	**C**	CATGAGTTACTA
MM 0 Allele **B**	ATCGGTAGCCAT	**G**	CATGAGTTACTA

Central probe quartet

Tiling strategy, 2



SNP **+4** Position

A / G

TAGCCATCGGTA **N** GTA **C** TCAATGATCAGCT

PM +4 Allele **A**	GTAGCCAT	T	CAT	**G**	AGTTACTAGTCG
MM +4 Allele **A**	GTAGCCAT	T	CAT	**C**	AGTTACTAGTCG
PM +4 Allele **B**	GTAGCCAT	C	CAT	**G**	AGTTACTAGTCG
MM +4 Allele **B**	GTAGCCAT	C	CAT	**C**	AGTTACTAGTCG

+4 offset probe quartet

Affymetrix SNP probe tiling strategy, 3

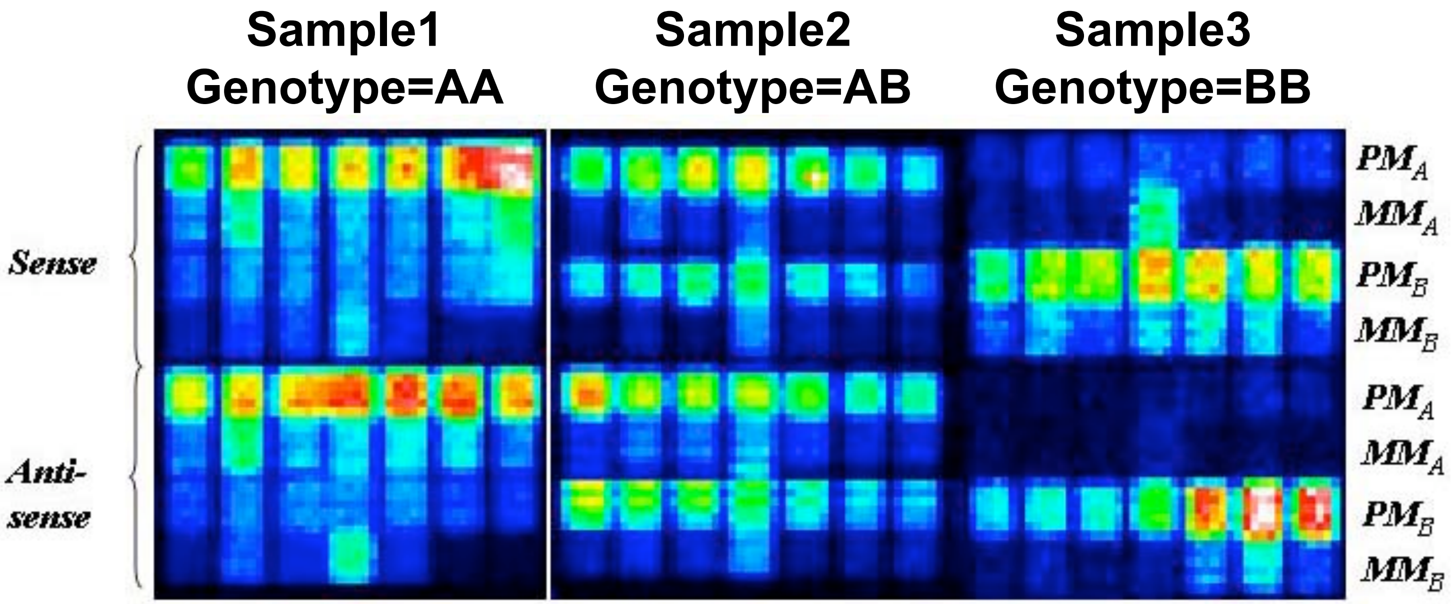
Offset quartets Central quartet Offset quartets

1	2	3	4	5	6	7
PMA	PMA	PMA	PMA	PMA	PMA	PMA
MMA	MMA	MMA	MMA	MMA	MMA	MMA
PMB	PMB	PMB	PMB	PMB	PMB	PMB
MMB	MMB	MMB	MMB	MMB	MMB	MMB

Repeated on the opposite strand: 56 probes for 10K.
More recently, 40: just 4 offset quartets instead of 6.

Probe Intensities

Fake (idealized) image for 3 samples on one SNP



Fake, as the probes are not all adjacent on the chip
Idealized, as all the probes are high or low as they should be.